НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра математичного моделювання та аналізу даних

«До захисту допущено»
В.о. завідувача кафедри
_____ Іван ТЕРЕЩЕНКО
«__» _____ 2023 р.

Дипломна робота

на здобуття ступеня бакалавра

зі спеціальності: 113 Прикладна математика
на тему: **«Математичні методи оцінки якості земельних ресурсів та прогнозування вартості землі»**

Виконав: студент 4 курсу, групи ФІ-91
Шередеко Олександр Володимирович

Керівник: доктор філософії з прикладної математики Яйлимова Г.О.

Консультант: _ _____

Рецензент: звання, степінь, посада Прізвище І.П. _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент _____

Київ — 2023

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра математичного моделювання та аналізу даних

Рівень вищої освіти — перший (бакалаврський)
Спеціальність (освітня програма) — 113 Прикладна математика,
ОПП «Кафедра математичного моделювання та аналізу даних»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Іван ТЕРЕЩЕНКО
«__» _____ 2023 р.

ЗАВДАННЯ
на дипломну роботу

Студент: Шередеко Олександр Володимирович

1. Тема роботи: «*Математичні методи оцінки якості земельних ресурсів та прогнозування вартості землі*»,

керівник:

доктор філософії з прикладної математики Яйлимова Г.О.,

затверджені наказом по університету №__ від «__» _____ 2023 р.

2. Термін подання студентом роботи: «__» _____ 2023 р.

3. Вихідні дані до роботи:

4. Зміст роботи: *огляд основних підходів до побудови регресійних та класифікаційних моделей, аналіз даних про якість та ціни землі, побудова та оцінка моделей регресії та класифікації*

5. Перелік ілюстративного матеріалу: *Презентація доповіді*

6. Дата видачі завдання: 26 січня 2023 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання	Примітка
1	Узгодження теми роботи із науковим керівником	15-30 січня 2023 р.	Виконано
2	Пошук джерела даних	01 - 15 лютого 2023 р.	Виконано
3	Дослідження предметної області	16 лютого - 31 березня 2023 р.	Виконано
4	Попередня обробка і аналіз даних	01 - 23 квітня 2023 р.	Виконано
5	Вивчення основних підходів до побудови регресії	24 квітня - 05 травня 2023 р.	Виконано
6	Побудова моделей та оцінка їх якості	08 - 31 травня 2023 р.	Виконано
7	Оформлення дипломної роботи	01 - 09 червня 2023 р.	Виконано
8	Отримання допуску до захисту	11 травня 2023 р.	Виконано

Студент _____ Шередеко О.В.

Керівник _____ Яйлимова Г.О.

РЕФЕРАТ

Кваліфікаційна робота містить: 69 стор., 21 рисунки, 52 таблиць, 10 джерел.

У даній роботі досліджуються дані про якість та ціну земель в Україні. Метою дослідження є аналіз якості та побудова прогнозу ціни землі використовуючи різні фактори. Для цього використовувалися математичні методи аналізу, моделі регресії та моделі класифікації. В ході дослідження було розглянуто основні підходи до методів математичного аналізу, до побудови регресійних моделей, зокрема лінійної,поліноміальної,лассо та дерев з градієнтним підсиленням, а також методи класифікації. Було виявлено основні залежності між ознаками у наборі даних, здіснено попередню обробку даних та протестовано декілька моделей, щоб знайти ту, яка працює найкраще. За допомогою моделей був побудований прогноз ціни землі та класифікація по різним факторам.

МАТЕМАТЕМАТИЧНІ МЕТОДИ АНАЛІЗУ, ВІЗУАЛІЗАЦІЯ , РЕГРЕСІЯ, КЛАСИФІКАЦІЯ

ABSTRACT

Qualification work contains: 69 pages, 21 figures, 52 tables, 10 sources.

This work analyzes data on the quality and price of land in Ukraine. The purpose of the study is to analyze the quality and build a forecast of land prices using various factors. For this purpose, mathematical methods of analysis, regression models and classification models were used. In the course of the study, the main approaches to mathematical analysis methods, to the construction of regression models, in particular linear, polynomial, lasso and gradient-boosted trees, as well as classification methods were considered. The main dependencies between the features in the dataset were identified, data preprocessing was performed, and several models were tested to find the one that workes best. The models were used to predict the price of land and classify it by various factors.

MATHEMATICAL METHODS OF ANALYSIS, VISUALIZATION, REGRESSION, CLASSIFICATION

ЗМІСТ

ВСТУП

Актуальність дослідження. Економіка України значною мірою залежить від сільського господарства, тому для фермерів та агробізнесу вкрай важливо мати доступ до прогнозів якості та ціни землі. Ця інформація дозволяє їм приймати обґрунтовані рішення щодо придбання землі, планування посівів та розподілу ресурсів. Оцінки якості та ціни землі можуть допомогти державним органам визначити території з високим потенціалом розвитку, а також ті, що потребують збереження з екологічних причин. Розуміючи фактори, які впливають на якість землі, інвестори можуть приймати стратегічні рішення та визначати потенційно вигідні інвестиційні можливості на українському ринку землі. Поточний, 2023 рік, робить це дослідження ще більш актуальним, прогнозування якості та ціни землі може відіграти важливу роль у відновленні економіки України, сприяючи зростанню в ключових секторах.

Метою дослідження є аналіз якості та побудова прогнозу ціни землі використовуючи різні фактори. Для досягнення цієї мети необхідно вирішити наступні завдання:

1) Вивчити основні підходи до побудови регресійних моделей;

2) Проаналізувати надані дані та виконати попередню обробку даних;

3) Провести аналіз даних за допомогою статистичних показників та виконати розвідувальний аналіз даних;

4) Підібрати та побудувати моделі, оцінити їх ефективність.

Об'єктом дослідження є дані, що стосуються якості та ціни землі в Україні.

Предметом дослідження є регресійні моделі для прогнозування якості та ціни землі.

При розв'язанні поставлених завдань використовувались такі *методи дослідження*: методи машинного навчання, методи математичної

статистики.

Практичне значення полягає в тому, що воно здатне надати інформацію для прийняття рішень зацікавленими сторонами, що в кінцевому підсумку призведе до більш ефективного землекористування, розподілу ресурсів та сталого розвитку.

Апробація результатів та публікації. Результати дослідження частково були представлені на XXII Науково-практичній конференції студентів, аспірантів та молодих вчених «Теоретичні і прикладні проблеми фізики, математики та інформатики» (11 травня 2023р., м. Київ).

1 АНАЛІЗ ДАНИХ ЗА ДОПОМОГОЮ СТАТИСТИЧНИХ ПОКАЗНИКІВ ТА ДОСЛІДНИЦЬКИЙ АНАЛІЗ ДАНИХ

1.1 Аналіз даних за допомогою статистичних показників

У цьому розділі буде проведений комплексний статистичний аналіз набору даних, використовуючи різні статистичні показники, такі як середнє значення, стандартне відхилення, дисперсія, асиметрія та коефіцієнт ексцесу. Ці показники допоможуть зрозуміти основні закономірності та розподіли даних, що матиме вирішальне значення для побудови моделі прогнозування якості та ціни землі.

Середнє значення [1] - міра центральної тенденції, яка представляє середнє значення набору чисел. Середнє значення може бути корисним представленням набору даних, коли дані приблизно симетричні і не мають екстремальних відхилень.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Дисперсія [1] - міра розсіювання, яка кількісно виражає відхилення окремих точок від середнього значення. Вона допомагає зрозуміти ступінь варіації.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Середньоквадратичне відхилення[1] використовується для оцінки розкиду даних і визначення ступеня варіації в наборі даних. Мале стандартне відхилення вказує на те, що точки даних близькі до середнього значення, тоді як велике стандартне відхилення вказує на те, що точки даних більш розкидані.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Асиметрія [2] - міра асиметрії або відсутності симетрії в розподілі набору даних. Це спосіб кількісно оцінити, наскільки розподіл точок даних відхиляється від ідеально симетричної, відомої як нормальний розподіл. Асиметрія допомагає визначити, чи є точки даних більш сконцентрованими

з одного боку від середнього значення порівняно з іншим.

$$\gamma = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)\sigma^3}$$

Значення асиметрії можна інтерпретувати наступним чином:

– Якщо асиметрія = 0, то розподіл симетричний навколо середнього значення.

– Якщо асиметрія > 0, розподіл є позитивно асиметричним або правостороннім, що означає, що точки даних більше сконцентровані з лівого боку від середнього значення, а з правого боку є довший хвіст.

– Якщо асиметрія < 0, розподіл є негативно асиметричним або лівостороннім, що означає, що точки даних більше зосереджені з правого боку від середнього значення, а з лівого боку є довший хвіст.

Коефіцієнт ексцессу [2] - міра, яка описує форму розподілу ймовірностей, зокрема, з точки зору його хвостів і піків. Він дає уявлення про хвостатість і концентрацію точок даних навколо середнього значення. Коефіцієнт ексцессу використовується разом з асиметрією, щоб надати більш повне розуміння форми розподілу.

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n\sigma^4}$$

Значення коефіцієнту ексцессу можна інтерпретувати наступним чином:

– Якщо коефіцієнт ексцессу = 0, то розподіл має такий самий коефіцієнт ексцессу, як і нормальний розподіл (мезокуртичний).

– Якщо коефіцієнт ексцессу > 0, то розподіл має більш важкі хвости і більш загострену середину порівняно з нормальним розподілом. Це означає, що існує більша ймовірність екстремальних значень у наборі даних.

– Якщо коефіцієнт ексцессу < 0, розподіл має легші хвости і пласкіший пік. Це означає, що екстремальні значення менш ймовірні в наборі даних.

Я провів аналіз даних щодо ціни за гектар у різних регіонах, метою було знайти середню ціну за гектар та відповідне стандартне відхилення. Отримані результати представлені у таблицях 1.1 - 1.5 нижче :

Таблиця 1.1 – Західна Україна

Oblast	Mean price per hectare	Standard deviation
Львівська	12928	13525
Тернопільська	22629	14727
Івано-Франківська	13584	14970
Волинська	16720	11337
Рівненська	11121	11784
Хмельницька	22657	14644
Чернівецька	25933	19676
Закарпатська	19139	14547

Таблиця 1.2 – Центральна Україна

Oblast	Mean price per hectare	Standard deviation
Вінницька	21583	13929
Дніпропетровська	25519	10892
Кіровоградська	27460	10060
Полтавська	24511	13251
Черкаська	25976	14083

Таблиця 1.3 – Північна Україна

Oblast	Mean price per hectare	Standard deviation
Житомирська	16181	12430
Київська	16463	14608
Чернігівська	17058	10635
Сумська	17652	12232

Таблиця 1.4 – Південна Україна

Oblast	Mean price per hectare	Standard deviation
Запорізька	20574	16630
Херсонська	22249	6668
Одеська	25636	10715

Таблиця 1.5 – Східна Україна

Oblast	Mean price per hectare	Standard deviation
Харківська	27635	10450
Донецька	24281	11284
Луганська	24700	10497

Центральний та східний регіони мають тенденцію до вищих середніх цін за гектар порівняно із західними та північними регіонами. Це може свідчити про такі фактори, як краща якість ґрунту, доступність інфраструктури або близькість до міських центрів, які роблять землю в цих регіонах більш цінною. Високе середньоквадратичне відхилення свідчить про широкий діапазон цін, що, можливо, пов'язано з різним набором характеристик землі, таких як якість ґрунту, топографія або землекористування, в межах регіону. І навпаки, низьке середньоквадратичне відхилення вказує на більшу узгодженість цін на землю, що може бути результатом більш однорідних характеристик землі або схожих ринкових умов.

Для подальшого аналізу, щоб краще зрозуміти нюанси здоров'я посівів на досліджуваній території, важливо заглибитися в дисперсію та середнє значення NDVI(оцінка щільності рослинності) за роками, що продемонстровано у таблиці 1.6 :

Таблиця 1.6 – Оцінка щільності рослинності 2019 - 2021

NDVI	Mean NDVI	Varience
NDVI 2019	6.15	0.55
NDVI 2020	6.1	0.66
NDVI 2021	5.51	0.43

Значення NDVI демонструють незначну тенденцію до зниження. Дисперсія, яка вимірює розкид набору даних є найвищою у 2020 році, що свідчить про те, що значення NDVI у цьому році були більш варіабельними порівняно з іншими роками. Це свідчить про те, що у 2020 році спостерігався вищий рівень мінливості стану посівів або густоти рослинності.

Щоб отримати більш повну картину, давайте заглибимося в характеристику якості землі - Бонітет, дослідивши його коефіцієнт ексцессу у різних регіонах. Аналіз коефіцієнту ексцессу для змінної бонітету у таблицях 1.7 - 1.11 в різних регіонах дає змогу отримати уявлення про розподіл якості земель в Україні.

Таблиця 1.7 – Західна Україна

Oblast	Bonitet Kurtosis
Львівська	7.891288415318289
Тернопільська	-0.07071763749042663
Івано-Франківська	-1.0441687481930388
Волинська	0.7492502953704688
Рівненська	-0.34530825491716505
Хмельницька	9.421250008462325
Чернівецька	15.529784184358002
Закарпатська	26.39423186249139

Таблиця 1.8 – Центральна Україна

Oblast	Bonitet Kurtosis
Вінницька	8.366586284543393
Дніпропетровська	26.745244962666156
Кіровоградська	28.67725701507058
Полтавська	8.90185428834715
Черкаська	12.051490266407942

Таблиця 1.9 – Північна Україна

Oblast	Bonitet Kurtosis
Житомирська	-0.8763696159131685
Київська	2.9557131120447764
Чернігівська	-0.9961644267570349
Сумська	4.301425457333816

Таблиця 1.10 – Південна Україна

Oblast	Bonitet Kurtosis
Запорізька	27.54727360110995
Херсонська	40.328607083311574
Одеська	33.99881751466644

Таблиця 1.11 – Східна Україна

Oblast	Bonitet Kurtosis
Харківська	26.25892709413327
Донецька	104.73153160839585
Луганська	84.00689859650566

У Західній Україні коефіцієнту ексцессу значно варіює між областями, що свідчить про різноманітні форми розподілу. Деякі області, такі як Львівська та Хмельницька, демонструють високий коефіцієнт ексцессу, що вказує на лептоподібний розподіл, який має екстремальні значення, що вказує на потенційні території з надзвичайно високою або низькою якістю земель. З іншого боку, від'ємний коефіцієнт ексцессу у таких областях, як Тернопільська та Івано-Франківська, вказує на платокуртичний розподіл, що означає ширший розподіл і менше екстремальних значень якості земель, а отже, більш рівномірну якість земель.

Подібні патерни з різним коефіцієнтом ексцессу можна спостерігати в Центральній, Північній та Південній Україні. Однак, варто відзначити високі значення коефіцієнту ексцессу на півдні та сході України, особливо в Донецькій та Луганській областях. Ці області демонструють надзвичайно високі значення коефіцієнту ексцессу, що свідчить про наявність значних відхилень у якості земель, які можуть бути зумовлені різними факторами, такими як різна родючість ґрунтів або практика землекористування.

Загалом, аналіз коефіцієнта ексцессу бонітета в різних регіонах України дає цінне розуміння розподілу якості земель в Україні. Він виявляє регіони з потенційними відхиленнями, які можуть стати об'єктом подальших спеціальних стратегій управління земельними ресурсами. Він також виділяє регіони з більш однорідною якістю земель, що може бути важливим для планування великомасштабної сільськогосподарської діяльності.

Аналіз асиметрії та середньої висоти над рівнем моря

продемонстрований у таблицях 1.12 - 1.16 в різних регіонах України дає змогу зробити деякі висновки щодо географічної структури країни. Асиметрія, яка є мірою розподілу ймовірностей, надає нам інформацію про те, як змінюється висота в межах країни.

Таблиця 1.12 – Західна Україна

Oblast	Skewness of Altitude	Mean Altitude
Львівська	3.5596502739802705	292.12469817418014
Тернопільська	-0.5374781713393317	316.95059006373776
Івано-Франківська	2.2934369245335877	357.3720650770948
Волинська	0.4445521429557412	198.94474205193478
Рівненська	0.15816267896956915	210.4029952773075
Хмельницька	-0.7980893761106911	288.7929541897143
Чернівецька	3.293966018498244	327.33921452231255
Закарпатська	2.8448966110313556	217.8274818663528

Таблиця 1.13 – Центральна Україна

Oblast	Skewness of Altitude	Mean Altitude
Вінницька	-1.194607997765679	257.9237014002567
Дніпропетровська	0.019272315687686937	106.06967817916139
Кіровоградська	-0.34797496905796715	166.90421915914052
Полтавська	0.1592794594380275	113.32230254357333
Черкаська	-0.20117977613156135	165.78391396927196

Таблиця 1.14 – Північна Україна

Oblast	Skewness of Altitude	Mean Altitude
Житомирська	-0.4024175792359955	218.96796797236712
Київська	0.262962031634871	146.34035938835237
Чернігівська	1.152891411855047	128.43854529872087
Сумська	-0.035540877405685	159.01832985115024

Таблиця 1.15 – Південна Україна

Oblast	Skewness of Altitude	Mean Altitude
Запорізька	0.5887221965493663	76.45071428539174
Херсонська	0.4022213984785889	38.16395551224059
Одеська	0.5425051820203945	100.65169757149167

Таблиця 1.16 – Східна Україна

Oblast	Skewness of Altitude	Mean Altitude
Харківська	-0.32476700796703895	152.5937725505011
Донецька	-0.7249479663966775	153.86358629303362
Луганська	-0.034539579415545016	124.12630601041398

У західних регіонах значення асиметрії значно відрізняються, що свідчить про різноманітний рельєф, з поєднанням рівнинних і горбистих регіонів. Це також підтверджується середніми значеннями висоти над рівнем моря, які є відносно високими. Високий показник асиметрії в деяких областях, таких як Львівська та Чернівецька, свідчить про значну кількість областей з більшою висотою над рівнем моря порівняно з рештою.

Центральна Україна демонструє відносно нижчі значення асиметрії та середньої висоти, що свідчить про більш збалансований та менш крутий рельєф. Північний регіон, з іншого боку, демонструє дещо більшу варіативність асиметрії, що свідчить про більш різноманітний рельєф з поєднанням високих і низьких ділянок.

Південна Україна демонструє позитивну асиметрію з низькими висотами над рівнем моря, що свідчить про те, що хоча регіон загалом рівнинний, є окремі райони з більшою висотою над рівнем моря. У Східній Україні асиметрія є від'ємною, що вказує на потенційну наявність більшої кількості територій з меншою висотою над рівнем моря.

У таблицях 1.17 - 1.21 наведено аналіз розподілу площ полів у різних регіонах України.

Таблиця 1.17 – Західна Україна

Oblast	Mean of areas	Std of areas	Variance of areas	Skewness of areas	Kurtosis of areas
Львівська	1.19	2.26	5.13	5.75	37.94
Тернопільська	1.72	2.11	4.46	9.79	146.3
Івано-Франківська	0.73	1.42	2.02	6.7	58.04
Волинська	1.54	2.03	4.15	12.62	244.1
Рівненська	1.18	1.11	1.25	1.59	5.24
Хмельницька	1.94	1.97	3.89	13.34	280.8
Чернівецька	1.05	1.00	1.01	4.12	39.7
Закарпатська	1.02	0.88	0.77	0.90	0.97

Таблиця 1.18 – Центральна Україна

Oblast	Mean of areas	Std of areas	Variance of areas	Skewness of areas	Kurtosis of areas
Вінницька	2.08	1.81	3.27	15.62	396.5
Дніпропетровська	4.43	4.51	20.4	6.8	78.31
Кіровоградська	3.43	2.42	5.89	2.76	26.05
Полтавська	2.66	1.89	3.59	5.40	104.9
Черкаська	2.16	1.77	2.95	4.73	40.05

Таблиця 1.19 – Північна Україна

Oblast	Mean of areas	Std of areas	Variance of areas	Skewness of areas	Kurtosis of areas
Житомирська	2.053	2.9	8.42	9.75	159.05
Київська	1.94	29.29	858.40	60.64	3720.7
Чернігівська	2.04	1.85	3.44	5.29	59.47
Сумська	2.07	3.76	13.74	20.22	525.72

Таблиця 1.20 – Південна Україна

Oblast	Mean of areas	Std of areas	Variance of areas	Skewness of areas	Kurtosis of areas
Запорізька	3.95	3.8	14.49	2.2	14.3
Херсонська	4.49	3.89	15.13	3.18	23.63
Одеська	2.87	3.06	9.4	2.21	7.32

Таблиця 1.21 – Східна Україна

Oblast	Mean of areas	Std of areas	Variance of areas	Skewness of areas	Kurtosis of areas
Харківська	3.84	2.86	8.17	3.27	33.6
Донецька	4.26	4.06	16.534	7.11	78.72
Луганська	3.29	3.54	12.58	7.24	139.69

Загалом, ці статистичні виміри вказують на значні відмінності в розмірах та розподілі сільськогосподарських полів, що відображають різноманітні регіональні практики, місцеві географічні особливості та історичні впливи.

Деякі регіони мають більші середні розміри полів, що свідчить про, можливо, більш екстенсивну сільськогосподарську практику, тоді як інші - менші, що вказує на менші розміри фермерських господарств. Західна Україна, що включає такі області, як Тернопільська та Волинська, має тенденцію до менших середніх розмірів полів. Це може свідчити про більш традиційні, дрібномасштабні методи ведення сільського господарства. На противагу цьому, такі регіони, як Вінниця, Дніпропетровщина, Харків та Донецьк мають більші середні розміри полів, що свідчить про наявність

великих сільських господарств.

Дисперсія розмірів полів, дає важливе уявлення про різноманітність методів ведення сільського господарства та форм власності на землю в регіоні. У Західній Україні деякі області, такі як Львівська та Тернопільська, демонструють більшу варіацію розмірів полів. Запорізька та Херсонська області, демонструють малу дисперсію, що свідчить про більшу однорідність у розмірах полів. Це може відображати більш диверсифікований сільськогосподарський сектор, який складається з малих, середніх та великих господарств.

Середньоквадратичне відхилення доповнює цю картину, вимірюючи дисперсію розмірів полів навколо середнього значення. Високе середньоквадратичне відхилення вказує на поєднання малих і великих фермерських господарств у регіоні. У Західній Україні деякі області, мають вищий показник стандартного відхилення, що вказує на більший розкид розмірів полів.

Позитивна асиметрія в багатьох регіонах свідчить про те, що малих полів більше, ніж великих, оскільки розподіл зміщений до нижньої межі. Області на півночі України, включаючи Київську та Чернігівську, демонструють позитивну асиметрію, що означає переважання малих та середніх господарств, з меншою кількістю великих господарств.

Високі значення коефіцієнту ексцессу в декількох регіонах свідчать про те, що існує більше екстремальних значень, ніж можна було б очікувати при нормальному розподілі. Вищий показник коефіцієнту ексцессу в деяких регіонах, таких як Вінниця та Дніпропетровщина, означає більше екстремальних значень. Ця особливість свідчить про складний і різноманітний сільськогосподарський ландшафт, з поєднанням малих фермерських господарств та великих сільськогосподарських підприємств.

Розуміння цих відмінностей має значення для сільськогосподарського планування та формування політики, забезпечуючи ефективне спрямування ресурсів та заходів на потреби

різних фермерських громад.

Заглиблюючись у дослідження статистичних індикаторів, я вирішив використовувати перехресну табуляцію, парні t-тести та ANOVA, що в сукупності допоможе краще проаналізувати дані. Також це допоможе в виборі данних для подальшого аналізу регресій, оскільки ми дізнаємося чи існує взаємозв'язок між даними.

Метод перехресної табуляції [3] дає загальне уявлення про взаємозв'язок між двома змінними і може допомогти виявити закономірності, якщо такі є, між змінними. Це можна проілюструвати за допомогою таблиці, яку ми отримуємо після дослідження, яка включає у себе значення Хі-квадрат, Хі-квадрат(DF), Хі-квадрат(Prop).

Хі-квадрат [4] - міра різниці між спостережуваними значеннями в категоріях перехресної таблиці та значеннями, які ми могли б очікувати отримати виключно випадково. Іншими словами, він показує, наскільки спостережувані дані відхиляються від нульової гіпотези про відсутність зв'язку між змінними. Високе значення хі-квадрат вказує на те, що змінні не є незалежними і мають значний зв'язок.

Хі-квадрат (DF) - кількість значень, які можуть варіюватися. У тесті хі-квадрат ступені свободи обчислюються як $(i-1)*(j-1)$.

Хі-квадрат(Prop) - ймовірність того, що нульова гіпотеза вірна. Якщо p-значення невелике, зазвичай менше 0,05, ми відкидаємо нульову гіпотезу і робимо висновок про існування зв'язку між змінними.

Аналіз методом перехресної табуляції допомагає зрозуміти зв'язок між двома змінними, а також те, чи є зв'язок, що спостерігається в даних, випадковим (що є нульовою гіпотезою), або ж він є статистично значущим.

Тест хі-квадрат демонструє значну кореляцію між класом земного покриву у 2021 році та основною вирощуваною культурою за період з 2016 по 2021 рік, що продемонстровано у таблиці 1.22 :

Таблиця 1.22 – Перехресна табуляція 1

Var1	Var2	Chi-square	Chi-square(DF)	Chi-square(Prop)
LC 2021	MJR CR	22940	384	0.628

Значення хі-квадрат і ступінь свободи, призводить до p-значення, близького до 0, свідчить про значний зв'язок між цими змінними. Таке високе значення хі-квадрат означає, що вибір сільськогосподарських культур нерозривно пов'язаний з основною вирощуваною культурою, що відповідає унікальним агрономічним вимогам різних культур, включаючи характеристики ґрунту, кліматичні умови та інтенсивність освітлення.

Тест хі-квадрат, результати якого наведено у таблиці 1.23 також виявляє послідовний зв'язок між областю та основною вирощуваною культурою.

Таблиця 1.23 – Перехресна табуляція 2

Var1	Var2	Chi-square	Chi-square(DF)	Chi-square(Prop)
Oblast	MJR CR	141116	208	0.0274

Великі значення хі-квадрат та ступенів свободи ,призводить до майже нульового p-значення, очевидна значна кореляція між цими змінними. Регіональне планування та вибір сільськогосподарських культур нерозривно пов'язані між собою, що посилює важливість регіональної сільськогосподарської стратегії для оптимізації продуктивності.

Тест хі-квадрат, результати якого наведено у таблиці 1.24 не виявив зв'язоку між типом власності та основною вирощуваною культурою.

Таблиця 1.24 – Перехресна табуляція 3

Var1	Var2	Chi-square	Chi-square(DF)	Chi-square(Prop)
MJR CR	ftype	463	7	11.235

Тест хі-квадрат показує що не існує сильного зв'язоку між типом основної вирощуваної культури та типом власності. Різниця в сільськогосподарських практиках та цілях між комерційними та приватними господарствами може пояснювати це. Комерційні господарства можуть віддавати перевагу культурам з високим попитом відповідно до ринкових тенденцій, в той час як приватні господарства можуть сприяти вирощуванню різноманітних культур, що відображає особисті уподобання.

Парний t-тест [1] - статистична процедура, яка використовується для визначення того, чи дорівнює нулю середня різниця між двома наборами спостережень. У парному t-тесті кожен суб'єкт або об'єкт вимірюється двічі, в результаті чого утворюються пари спостережень.

Опис вихідних змінних парного t-тесту:

t - відношення відхилення оціненого значення параметра від його гіпотетичного значення до стандартної похибки.

Середнє значення стандартної похибки - оцінка мінливості між середнім значенням вибірки та середнім значенням генеральної сукупності. Вона дає нам міру точності нашої оцінки середнього значення генеральної сукупності.

Довірчий інтервал (нижня та верхня межі) - кінці діапазону значень середньої різниці, в яких ми можемо бути впевненими, що вони містять справжню середню різницю генеральної сукупності.

Парний t-тест надає інформацію про зв'язок між двома наборами парних спостережень, в тому числі про те, чи існує між ними значуща різниця, розмір і напрямок цієї різниці.

Парний t-тест, результати якого наведено у таблиці 1.25 для порівняння значень NDVI у 2019 та 2020 роках.

Таблиця 1.25 – Парний t-тест 1

Var1	Var2	t value	Std dev	STD error mean	Conf. interval lower	Conf interval upper
NDVI 2019	NDVI 2020	17.99	0.673	0.00295	0.047	0.059

Значення t-критерію виявилося суттєвим, що свідчить про помітну різницю між цими двома роками. Не менш показовою була мінімальна стандартна похибка середнього значення, яка вказує на те, що варіабельність навколо середнього значення була відносно низькою. Довірчий інтервал ще більше підкреслив значущість різниці між двома роками, чітко вкладаючись у вузький діапазон.

Парний t-тест, результати якого наведено у таблиці 1.26 для порівняння значень NDVI у 2020 та 2021 роках.

Таблиця 1.26 – Парний t-тест 2

Var1	Var2	t value	Std dev	STD error mean	Conf. interval lower	Conf interval upper
NDVI 2020	NDVI 2021	169.56	0.795	0.00348	0.584	0.0598

Високе t-значення, а також мале p-значення свідчать про те, що ця різниця є статистично значущою. Довірчий інтервал знаходиться в межах вузького діапазону, що посилює докази значної різниці між двома роками.

Парний t-тест, результати якого наведено у таблиці 1.27 для порівняння значень відстані до найближчого міста та села.

Таблиця 1.27 – Парний t-тест 3

Var1	Var2	t value	Std dev	STD error mean	Conf. interval lower	Conf interval upper
Dist Citie	Dist Villa	298.29	9.832	0.043	12.778	12.947

Високе t-значення і практично нульове p-значення вказують на суттєву і статистично значущу різницю. Вона свідчить про те, що фермерські господарства, як правило, розташовані ближче до сіл, ніж до міст, що може суттєво впливати на ланцюги поставок сільськогосподарської продукції та доступ до ринку.

Дисперсійний аналіз (ANOVA) [1] - це статистичний метод, який використовується для перевірки відмінностей між двома або більше середніми. ANOVA перевіряє гіпотезу про те, що середні значення

декількох популяцій є рівними. Нульова гіпотеза стверджує, що всі середні популяції рівні, тоді як альтернативна гіпотеза стверджує, що принаймні одна з них відрізняється.

Опис вихідних змінних ANOVA:

Сума квадратів - показник загальної мінливості даних. "Міжгрупова"сума квадратів представляє загальну мінливість, що пояснюється груповими відмінностями, а "внутрішньогрупова"сума квадратів представляє загальну мінливість, зумовлену відмінностями в межах кожної групи.

F-статистика - відношення середнього квадрата між групами до середнього квадрата всередині групи. Чим більша F-статистика, тим більше доказів проти нульової гіпотези про рівність середніх значень у генеральній сукупності.

Тест Левена використовується для перевірки припущення про однорідність дисперсій, яке є припущенням ANOVA.

ANOVA-тест, результати якого наведено у таблиці 1.28 був проведений для того, щоб з'ясувати, чи існують значні відмінності в якості землі між різними основними культурами.

Таблиця 1.28 – ANOVA 1

Var1	Var2	BG SS	WG SS	BG Mean	WG Mean	F	Levene's stat
Bonitet	MJR CR	377661	3427642	25177.437594	66.288429948	379.81647	124.5985

Результати тесту продемонстрували значну F-статистику. Таке велике значення F свідчить про значні відмінності між досліджуваними групами, причому міжгрупові відмінності значно перевищують внутрішньогрупові. Цей результат означає, що якість землі варіюється більш суттєво між різними типами сільськогосподарських культур, ніж у межах груп однієї культури. На додаток до цього, тест Левена, який оцінює однорідність дисперсій, показав, що дисперсії різних груп не є рівними. Це підкріплює висновок про те, що існують значні відмінності в

якості землі між різними типами сільськогосподарських культур.

Проведений тест ANOVA, результати якого наведено у таблиці 1.29 мав на меті визначити, чи існують значні відмінності в середній відстані до найближчого міста між різними регіонами.

Таблиця 1.29 – ANOVA 2

Var1	Var2	BG SS	WG SS	BG Mean	WG Mean	F	Levene's stat
Dist Citie	oblast	699695.35937	4362973.9137	29153.9733	83.9485475	347.283832	212.07863370

Тест показав велику F-статистику, що означає значну різницю між середніми значеннями різних груп. Міжгрупові відмінності перевищили внутрішньогрупові, що свідчить про те, що відстань до міст більше варіюється між різними регіонами, ніж у межах одного регіону. Крім того, показав, що дисперсії в різних регіонах не однакові, що свідчить про їхню неоднорідність. Відстань до міст - фактор, який може впливати на розподіл та управління сільськогосподарськими ресурсами - суттєво відрізняється в різних регіонах.

ANOVA тест, результати якого наведено у таблиці 1.30 був проведений для того, щоб перевірити, чи відрізняється середня площа поля між комерційною та приватною власністю.

Таблиця 1.30 – ANOVA 3

Var1	Var2	BG SS	WG SS	BG Mean	WG Mean	F	Levene's stat
area	ftype	14607.3410391	3357275.1683	14607.341	171.92990056	84.961027669	22.07761643

Результати тесту F-статистики свідчать про суттєві відмінності в середніх значеннях груп. Сума квадратів, як міжгрупових, так і внутрішньогрупових, також підтверджує цей результат. Також було проведено тест Левена, який показав нерівномірність дисперсій. Ці результати підкреслюють, що розміри полів в аграрному секторі суттєво відрізняються залежно від форми власності - комерційної чи приватної.

1.2 Розвідувальний аналіз даних

Візуалізація, за своєю суттю, є невід'ємною частиною аналізу даних, перетворюючи сирі, часто незбагненні цифри на зрозумілі, інтерпретовані і, перш за все, відчутні зображення даних. Візуалізація допоможе дослідити безліч взаємозв'язків, закономірностей і тенденцій. У наступному розділі використано низку методів візуалізації для вивчення різних аспектів даних. За допомогою візуалізації завданням є розкрити глибше розуміння даних, надаючи більш повне уявлення про різні фактори.

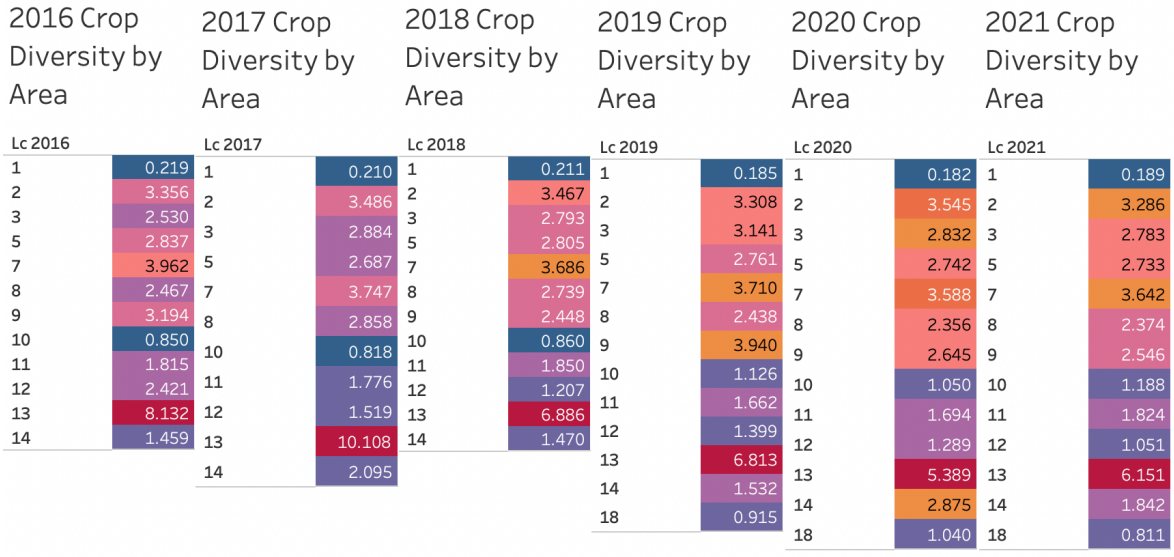


Рисунок 1.1 – Розподіл посівів та земних покривів за роками

Проаналізувавши теплові карти за різні роки, які продемонстровано на рисунку 1.1 можна помітити деякі важливі тенденції та цікаві факти.

З 2016 по 2021 рік середня площа поля для класу земного покриву, представленого ідентифікатором 13, що є водним покривом, постійно підтримувала більшу середню площу порівняно з іншими класами. На противагу цьому, клас типу земного покриву з ідентифікатором 1, що є штучним об'єктом постійно мав найменшу середню площу поля впродовж усіх років. Теплиці - це ефективний спосіб контролювати середовище вирощування рослин, що дозволяє вирощувати їх цілий рік незалежно від

зовнішніх погодних умов. Важливо, що вони не потребують великої площі, щоб бути ефективними. Таким чином, теплиці пропонують продуктивне використання землі, навіть коли простір обмежений.

Дивлячись на клас з ідентифікатором 2 що є крупами, ми бачимо незначне зменшення середньої площі з плином часу. Ця зміна може свідчити про поступовий перехід до більш інтенсивних методів ведення сільського господарства або про зміну культур, що вирощуються.

Крім того, дані показують постійну присутність певних класів таких як 3(ріпак), 5(кукурудза), 7(соняшник), 8(соя), 11(пасовище) і 14(болото) з відносно великими середніми площами полів протягом багатьох років, що свідчить про те, що ці типи ґрунтового покриву або сільськогосподарських культур залишаються стабільними і важливими у землекористуванні протягом тривалого часу.

Загалом, візуалізація теплової карти ефективно висвітлює динаміку та зміни в земному покриві та типах сільськогосподарських культур протягом багатьох років.

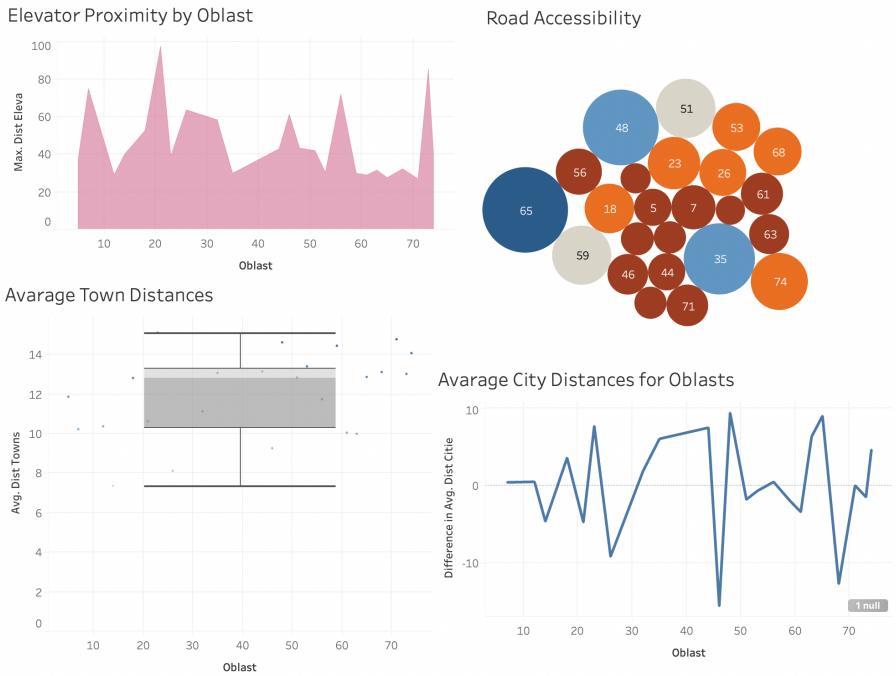


Рисунок 1.2 – Логістика

Давайте заглибимося в аналіз кожної з візуалізацій на рисунку 1.2 :

– 1. Області з максимальною відстанню до найближчого елеватора - це області 21(Закарпацька) та 73(Чернівецька), з відстанню приблизно 97,39 та 85,20 відповідно. Це означає, що ці регіони є найвіддаленішими від інфраструктури, яка допомагає у транспортуванні сільськогосподарської продукції, можливо, через їх географічне розташування або недостатній рівень розвитку. З іншого боку, області 71(Черкаська) та 65(Херсонська) мають найменшу максимальну відстань до найближчого елеватора, приблизно 26,80 та 27,51 км відповідно, що свідчить про більш доступну інфраструктуру в цих регіонах. Це надає можливість більш ефективному транспортуванню та дистрибуції.

– Область з найвищою медіанною відстанню до найближчої сільської дороги - 65(Херсонська), що становить приблизно 2,45. Це свідчить про те, що центральна тенденція відстані до сільських доріг у цьому регіоні є відносно високою, що вказує на потенційні транспортні проблеми в цьому регіоні через менш щільну мережу сільських доріг. Це може створити труднощі у сполученні та доступності для сільської місцевості. На противагу цьому, область 32(Київська) має найменшу середню відстань до найближчої сільської дороги - 0,286, що свідчить про більш густонаселену мережу сільських доріг. Це може означати більш ефективне транспортне сполучення та легший доступ до різних населених пунктів регіону, що принесе користь як мешканцям, так і підприємствам, які там працюють.

– Область з найбільшою середньою відстанню до найближчого міста - 23(Запоріжжя), приблизно 15,09 км. Такий високий середній показник свідчить про те, що більша частина території області знаходиться відносно далеко від найближчого міста, що вказує на те, що цей регіон може бути більш ізольованим. Наслідки цього можуть бути різними: від меншої доступності міських зручностей до потенційних проблем з транспортуванням товарів і людей. З іншого боку, область 14(Донецька) має найменшу середню відстань до найближчого міста -

7,35 км. Це може свідчити про більш урбанізований регіон або про те, що він знаходиться в безпосередній близькості до кількох міст, що полегшує доступ до міських ресурсів, послуг та потенційних ринків збуту сільськогосподарської продукції.

– Різниця у середньому показнику дистанції до найближчого міста від попереднього показника: область з найбільш значним зменшенням середньої відстані до найближчого міста порівняно з попереднім показником - 26(Івано-Франківська) зі зміною на -9,0792. Таке суттєве зменшення може свідчити про декілька речей. Наприклад, це може свідчити про значне зниження темпів розвитку міст або, можливо, про міграцію населення з міських територій у цьому регіоні. Це також може вказувати на зміну меж міст або створення нових міст ближче до цієї області. І навпаки, область 23(Запорізька) показала найбільш значне збільшення середньої відстані до найближчого міста, з різницею в 7,7084.

Ці вимірювання надають дані про географічні та інфраструктурні характеристики кожної області. Наприклад, райони з великими відстанями до елеваторів або міст можуть потребувати інвестицій в інфраструктуру для стимулювання економічної активності. З іншого боку, регіони з меншими відстанями можуть бути ідеальними для видів діяльності, які потребують частих поїздок або транспорту.

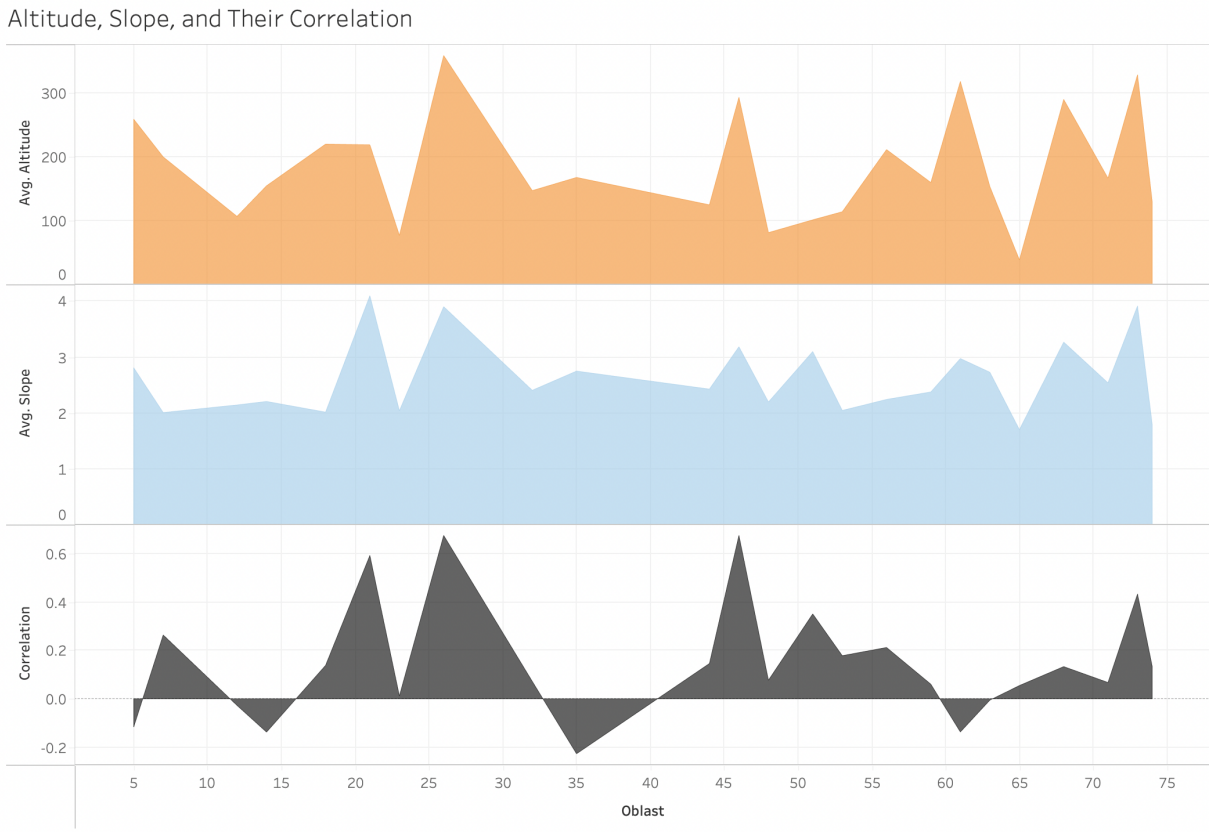


Рисунок 1.3 – Кореляція висоти та схилу

Аналізуючи візуалізацію подвійної комбінації для середньої висоти та середнього схилу в кожній області на рисунку 1.3, ми можемо спостерігати кореляції та тенденції в даних:

– Позитивна кореляція: для деяких областей спостерігається позитивна кореляція між середньою висотою та середнім схилом. Наприклад, області 21(Закарпаття), 26(Івано-Франківськ) та 46(Львів) демонструють відносно сильну позитивну кореляцію з коефіцієнтами кореляції 0,5908, 0,6730 та 0,6732 відповідно. Це вказує на те, що зі збільшенням середньої висоти над рівнем моря в цих областях середній нахил також має тенденцію до збільшення. Це може означати, що ці регіони мають більш гірський або горбистий ландшафт, що може вплинути на типи культур, які можна вирощувати та вимоги до інфраструктури цих територій. Ці регіони можуть потребувати спеціальних стратегій розвитку сільського господарства та інфраструктури з урахуванням рельєфу місцевості.

– Негативна кореляція: декілька областей демонструють негативну кореляцію між середньою висотою над рівнем моря та середнім схилом, наприклад, області 5(Вінницька), 12(Дніпропетровська), 14(Донецька), 35(Кіровоградська) та 61(Тернопільська), з коефіцієнтами кореляції -0,1167, -0,0275, -0,1376, -0,2268 та -0,1370 відповідно. Це означає, що зі збільшенням середньої висоти середній нахил має тенденцію до зменшення в цих регіонах. Це може свідчити про те, що ці райони мають змішаний ландшафт, з рівнинними та горбистими ділянками, що призводить до різноманітного рельєфу.

– Слабка кореляція або її відсутність: деякі області демонструють слабку кореляцію або її відсутність між середньою висотою над рівнем моря та середнім схилом, наприклад, області 23(Запорізька), 32(Київська), 53(Полтавська), 59(Сумська), 63(Харвіська), 65(Херсонська), з коефіцієнтами кореляції в діапазоні від -0,0057 до 0,1775. Це означає, що в цих регіонах може не існувати значного зв'язку між середньою висотою і схилом, або що на характеристики рельєфу впливають інші фактори. .

– Найвища та найнижча середня висота над рівнем моря: область з найвищою середньою висотою над рівнем моря - Чернівецька, з висотою приблизно 327,34, тоді як область з найнижчою середньою висотою над рівнем моря - Херсонська, з висотою приблизно 38,16. Така різниця у висоті над рівнем моря може впливати на такі фактори, як клімат, сільськогосподарські практики та вимоги до інфраструктури.

– Найвищий та найнижчий середній нахил: область з найбільшим середнім схилом - Запорізька, зі схилом приблизно 4,0930, тоді як область з найнижчим середнім схилом - Херсонська, з схилом приблизно 1,6970. Території з більшими схилами можуть потребувати специфічних сільськогосподарських практик, таких як терасування або вирощування культур, які можуть знаходитися на місцевості з більш крутим рельєфом.

Розуміння кореляції між висотою над рівнем моря та схилами в різних областях дає уявлення про географічні характеристики регіонів.

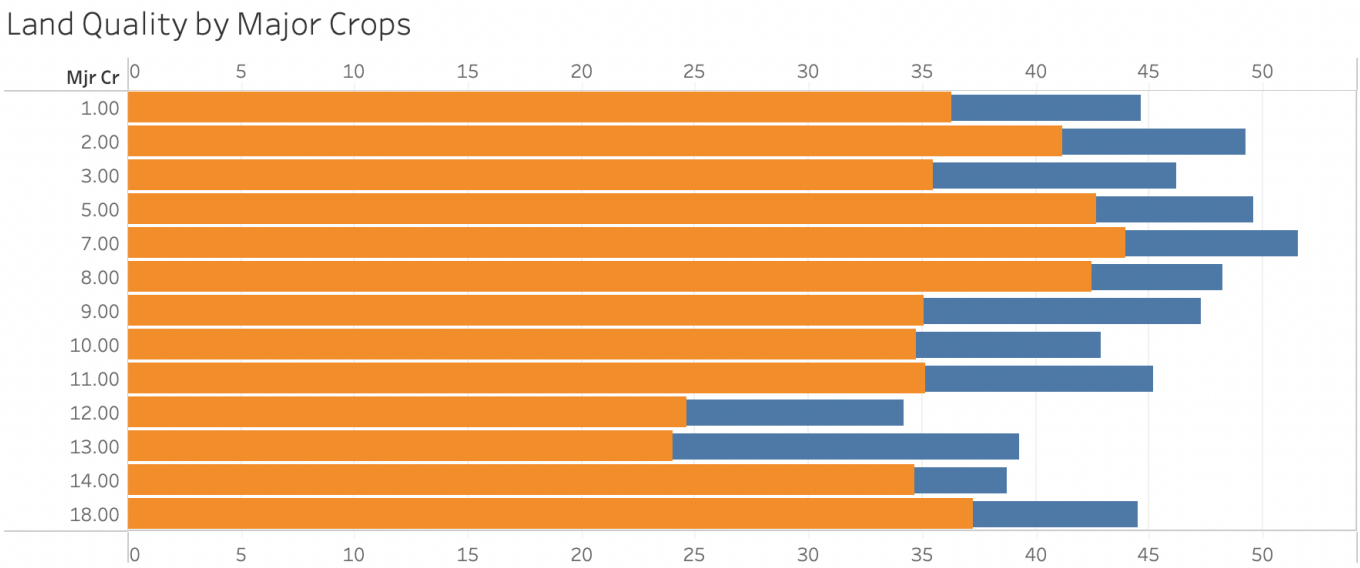


Рисунок 1.4 – Якість землі за основними вирощуваними культурами

Дивлячись на візуалізацію для середньої якості землі для кожної основної культури на рисунку 1.4, ми можемо зробити деякі цікаві висновки:

– Найвищий та найнижчий бонітет: основна культура соняшник має найвищу середню якість землі - 51,56, тоді як основний ґрунтовий покрив пустелі має найнижчу середню якість землі - 34,21. Різниця у значеннях бонітету означає, що якість землі суттєво відрізняється залежно від типу основної культури/типу ґрунтового покриву. Ця різниця в якості землі може впливати на врожайність сільськогосподарських культур, де вища якість землі, як правило, призводить до вищої врожайності.

– Основні культури такі як: соняшник, кукурудза, соя та крупа постійно демонструють вищі значення , що свідчить про те, що ці культури зазвичай вирощуються на високоякісних землях.

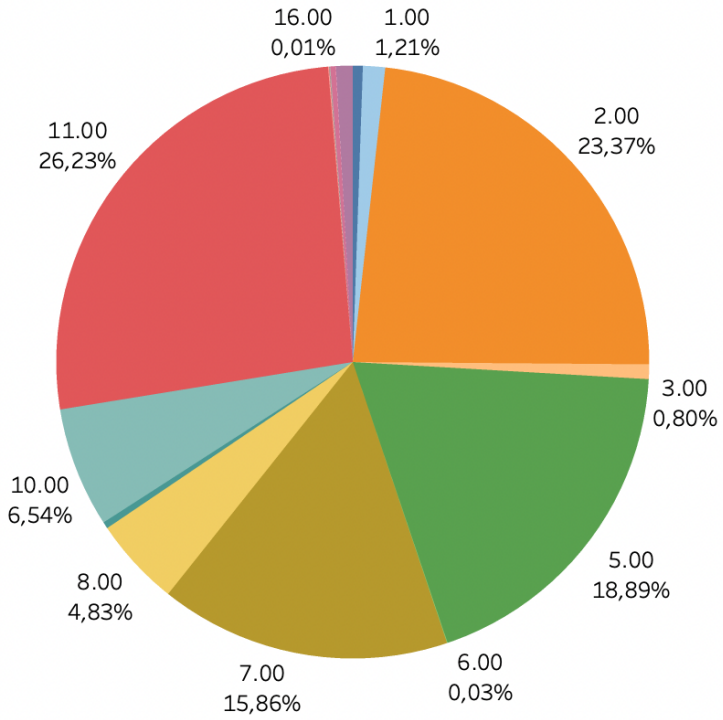


Рисунок 1.5 – Основні вирощувані культури за площею

Аналізуючи візуалізацію кругової діаграми розподілу основних сільськогосподарських культур у різних регіонах на рисунку 1.5, ми можемо зробити наступні спостереження:

– Найпоширеніші основні культури: основні культури круп та трав'яниста земля складають найбільш значний відсоток від загальної площі області - 23,3674 та 26,2283 відповідно. Це свідчить про те, що ці культури найчастіше вирощуються в усіх регіонах. Ці культури можуть бути основними сільськогосподарськими товарами в країні, роблячи значний внесок у економіку.

Таким чином, ці візуалізації надають інформацію про взаємозв'язок між основними сільськогосподарськими культурами та якістю землі, а також про розподіл основних сільськогосподарських культур в Україні. Ця інформація може допомогти у розробці сільськогосподарських стратегій та політик, спрямованих на оптимізацію врожайності, ефективне управління ресурсами.

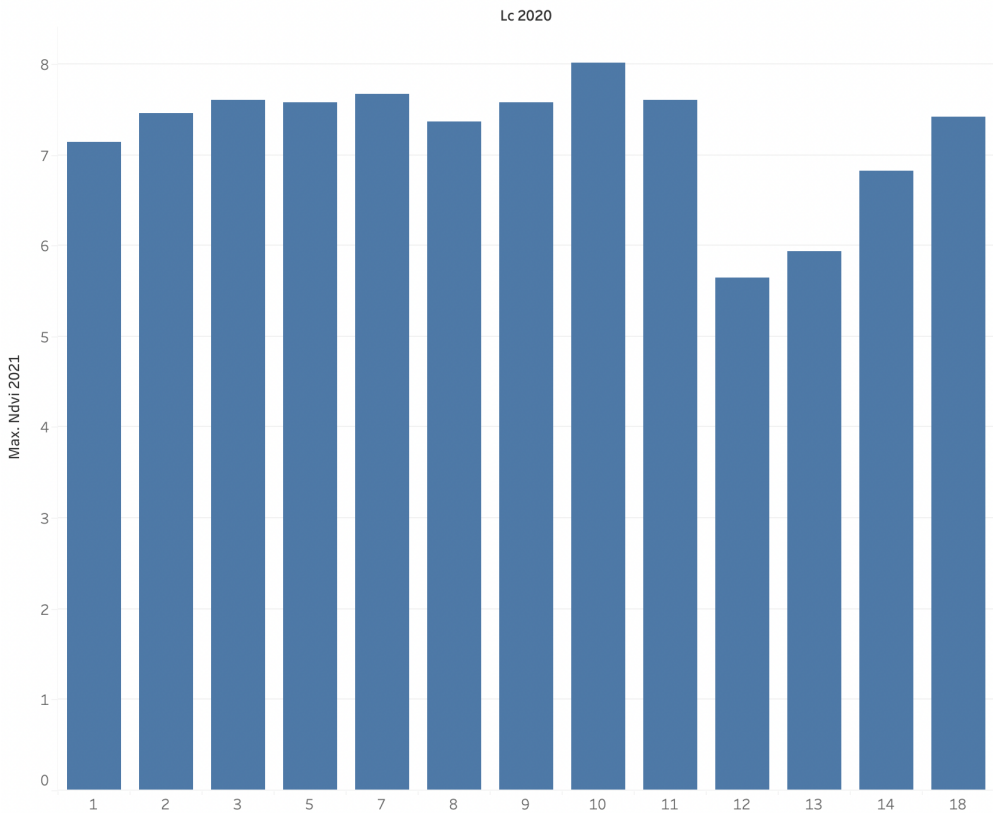


Рисунок 1.6 – Ґрунтовий покрив/тип сільськогосподарських культур 2020 -> NDVI 2021

Аналізуючи візуалізацію гістограми для максимального NDVI у 2021 році для кожного типу сільськогосподарських культур у 2020 році на рисунку 1.6, ми можемо зробити кілька висновків:

– Найвищий та найнижчий NDVI: тип культур лісу з 2020 року має найвище значення NDVI у 2021 році, з максимальним NDVI 8,009. Це означає, що ці ділянки мали найміцніший вегетативний стан у 2021 році, що може бути наслідком сприятливих умов вирощування або ефективних стратегій управління земельними ресурсами у попередньому році. І навпаки, ґрунтовий покрив болот з 2020 року має найнижчий NDVI у 2021 році - 5,6404. Ці ділянки продемонстрували найменший вегетативний індекс у 2021 році, що, можливо, пов'язано з менш сприятливими умовами або менш ефективним управлінням земельними ресурсами.

– Діапазон максимальних значень NDVI становить 8,009 - 5,6404 = 2,3686. Ця різниця вказує на значну варіацію вегетативного стану різних

типів культур, що підкреслює вплив вибору культури на подальший ріст рослинності.

– NDVI та тип культури: майже всі типи культур мають максимальне значення NDVI більше 7, що свідчить про те, що більшість культур у 2020 році забезпечили здоровий вегетативний індекс у 2021 році. Це свідчить про те, що широкий спектр культур може сприяти міцному здоров'ю рослин у наступні роки, підкреслюючи потенційні переваги сівозміни або диверсифікації.

– Вплив на майбутнє вегетативне здоров'я: тип культури, що вирощується в даному році, може мати значний вплив на вегетативний стан землі в наступному році. Наприклад, деякі культури можуть покращити стан ґрунту, поповнюючи його поживними речовинами, в той час як інші можуть виснажити ґрунт або підвищити сприйнятливість до шкідників чи хвороб. Ці висновки підкреслюють важливість продуманого сільськогосподарського планування та управління земельними ресурсами.

Таким чином, ця візуалізація дає розуміння взаємозв'язку між типом культури, вирощеної в одному році , та вегетативним здоров'ям землі у наступному році.

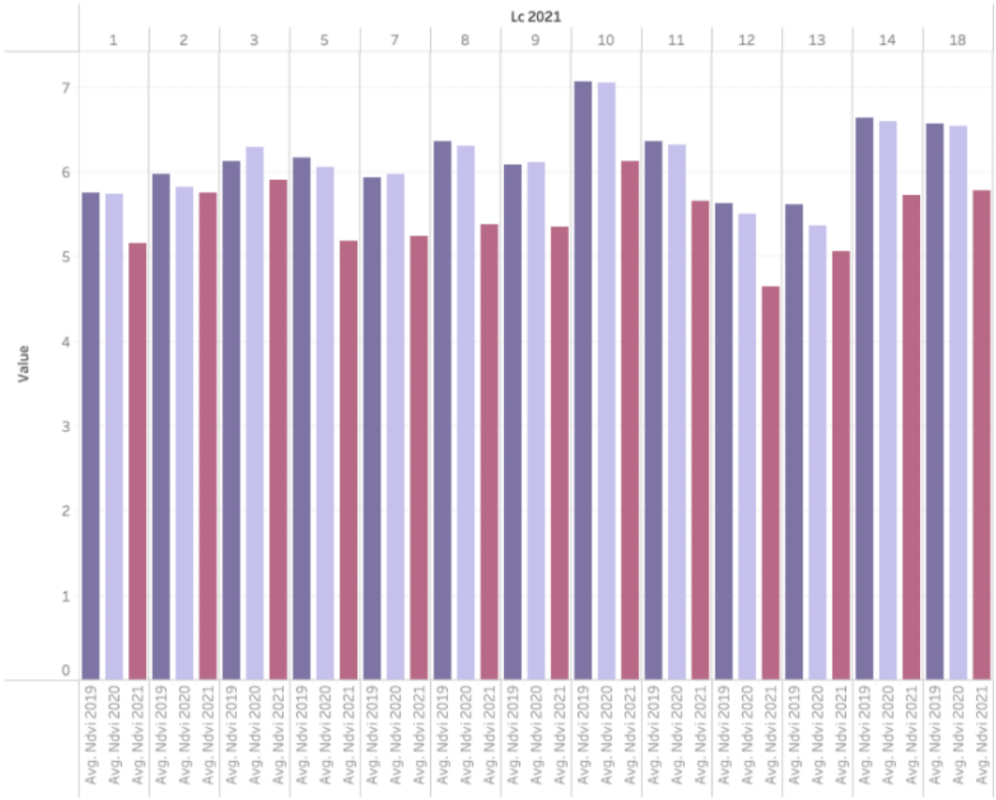


Рисунок 1.7 – Середні значення NDVI 2019-2021 та типи ґрунтового покриву/типу сільськогосподарських культур 2021

Стовпчаста гістограма на рисунку 1.7 надає детальне порівняння середніх значень NDVI для різних типів сільськогосподарських культур за три роки, з 2019 по 2021 рік. Розглянемо деякі спостереження:

– Тенденція в часі: для всіх типів сільськогосподарських культур, що входять до складу земельного покриву, спостерігається послідовне зниження середнього значення NDVI з 2019 по 2021 рік. Це може свідчити про тенденцію до погіршення вегетативного здоров'я протягом цих років, можливо, через зміну умов навколишнього середовища або сільськогосподарських практик.

– Зміна значень NDVI: величина зниження NDVI з 2019 по 2021 рік також варіюється між типами культур. Наприклад, NDVI для типу культур лісів зменшився з 7,07161 у 2019 році до 6,12423 у 2021 році, тобто на 0,94738. Це одне з найбільших падінь, що спостерігалося. На противагу цьому, NDVI для ґрунтового покрививі штучними об'єктами

зменшився лише з 5,75838 у 2019 році до 5,15759 у 2021 році, тобто на 0,60079, що є одним з найменших спостережуваних знижень.

Таким чином, візуалізація гістограми дає повне уявлення про середнє значення NDVI для кожного типу сільськогосподарських культур за три роки. Загальна тенденція до зниження NDVI може викликати занепокоєння, а різні темпи зниження серед різних типів культур свідчать про різний рівень їхньої стійкості або пристосованості до мінливих умов.

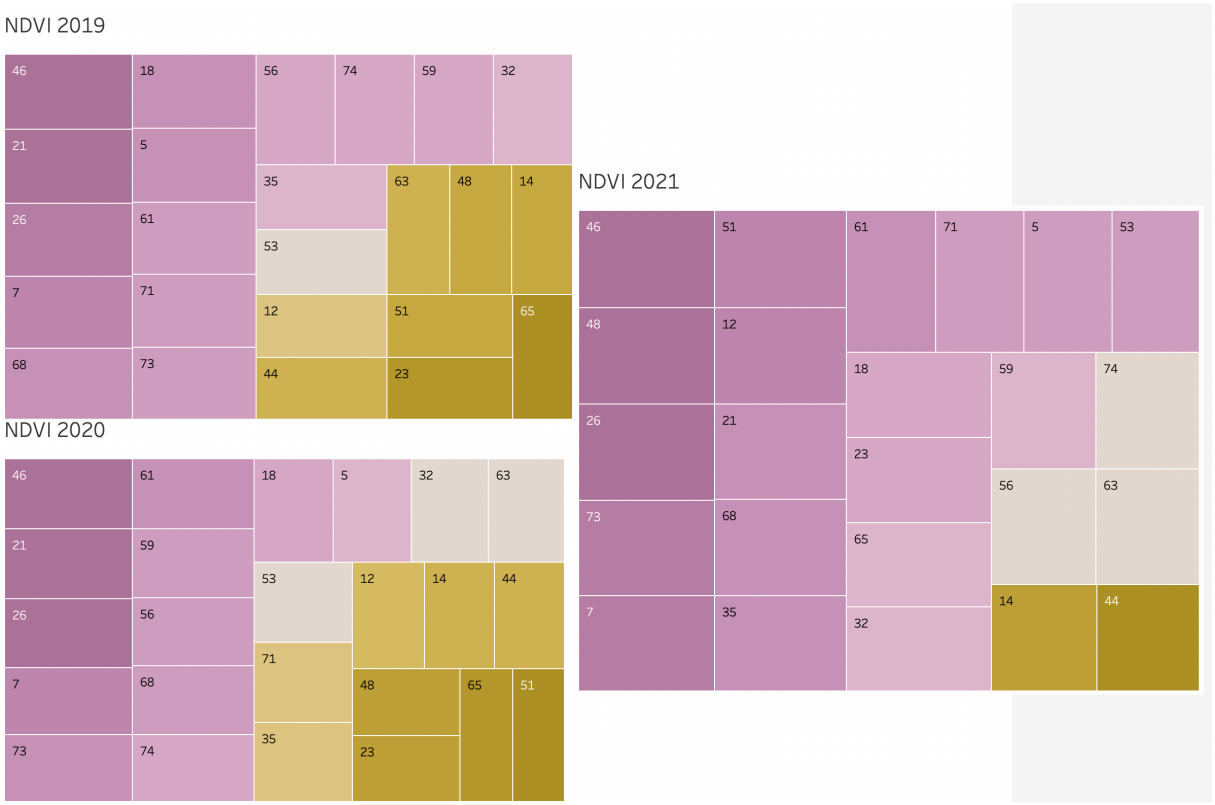


Рисунок 1.8 – NDVI у різних регіонах

Теплові карти на рисунку 1.8 показують середнє значення NDVI для різних регіонів за 2019, 2020 та 2021 роки.

– Загальний тренд: У середньому в більшості регіонів спостерігалося зниження NDVI з 2019 по 2021 рік, що свідчить про зменшення зеленого кольору рослинності та, можливо, здоров'я полів сільськогосподарських культур.

– Деякі області підтримували відносно високі значення NDVI протягом усіх трьох років. Наприклад, Львівська область постійно має

одне з найвищих значень NDVI, що свідчить про міцний стан рослинності протягом цього періоду. І навпаки, деякі області мають стабільно низькі значення NDVI. Зокрема, Луганська область має відносно низьке значення NDVI протягом усіх трьох років і демонструє значне зниження у 2021 році, що вказує на можливі проблеми зі станом рослинності в цьому регіоні.

– У деяких регіонах спостерігається значне зниження значень NDVI з плином часу. Наприклад, в Чернігівській спостерігається зниження з 6,26586 у 2019 році до 5,384223 у 2021 році. Хоча в більшості областей спостерігається зниження NDVI з роками, деякі з них продемонстрували незначне зростання. Зокрема, Одеська область зросла з 5,67421 у 2019 році до 5,701396 у 2021 році, незважаючи на падіння у 2020 році.

Таким чином, ці теплові карти надають візуальну ілюстрацію того, як значення NDVI змінювалися з часом у різних регіонах. Вони підкреслюють мінливість здоров'я та продуктивності рослинності в різних регіонах і з плином часу.

Висновки до розділу 1

У цьому розділі були використані статистичні інструменти для аналізу даних, що стосуються землі та її різноманітних характеристик. Застосовуючи статистику, були оцінені змінні, а також проаналізований звязок між ними. Аналіз дозволив встановити деякі цікаві тенденції та закономірності. Візуалізації покращили розуміння аналізу, представляючи висновки у більш доступному форматі.

2 РЕГРЕСІЙНИЙ ТА КЛАСИФІКАЦІЙНИЙ АНАЛІЗ

У наступному розділі дослідження ми перейдемо від описового аналізу до регресійного та класифікаційного аналізу. Регресійний аналіз є безцінним інструментом, який може дати уявлення про те, як багато факторів взаємодіють один з одним і як вони сукупно впливають на результат - ціну землі в Україні. Порівнюючи результати різних моделей, ми можемо визначити, яка з них є найкращою для нашого набору даних, і краще зрозуміти фактори, що впливають на ціни на землю в Україні. Класифікаційний аналіз дозволяє нам виявляти закономірності та взаємозв'язки в категоріальних даних. Алгоритми класифікації вміють працювати як з лінійними, так і з нелінійними зв'язками, що робить їх універсальними інструментами для різних даних. Це не тільки дасть нам краще розуміння поточного стану ринку землі, але й уможливить точніші прогнози на майбутнє, допомагаючи зацікавленим сторонам приймати обґрунтовані рішення.

2.1 Основні підходи до побудови регресії та класифікації

Лінійна регресія [5] - це статистичний метод, який використовується для вивчення зв'язку між двома неперервними змінними. Лінійна регресія знаходить зв'язок між залежною і незалежною змінними у вигляді прямої лінії, моделюючи залежну змінну як лінійну функцію незалежної змінної.

Лінійна регресія включає одну незалежну змінну і одну залежну змінну. Зв'язок між цими двома змінними описується наступним рівнянням:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Тут y_i позначає залежну змінну, x_i - незалежну змінну, β_0 - довжина відрізку від початку координат до точки перетину прямої з віссю OY, β_1 - тангенс кута нахилу прямої до осі OX, а ϵ_i є похибкою для i-го спостереження. У контексті статистичного моделювання коефіцієнти β_0 і β_1 вважаються параметрами популяції, тобто вони є істинними значеннями інтервалу і нахилу для всієї популяції. Оцінки $E(\beta_0)$ та $E(\beta_1)$ розраховуються за допомогою методу найменших квадратів, який мінімізує суму відхилень значень, що фактично спостерігаються, від модельних значень. Формули для них наступні:

$$E(\beta_1) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$E(\beta_0) = \bar{y} - E(\beta_1)\bar{x}$$

де $\bar{x}$ і $\bar{y}$ представляють середні значення x і y відповідно.

Для вимірювання мінливості відхилень значень ми використовуємо оцінку дисперсії вибірки, позначену як s^2, яка обчислюється наступним чином:

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

де n - кількість спостережень, а $\hat{y}_i$ - прогнозоване значення y_i на основі регресії.

Ми також можемо побудувати довірчий інтервал для β_1. Наприклад, 95 % довірчий інтервал можна побудувати наступним чином:

$$E(\beta_1) \pm t_{0.025,n-2} s_{b_1}$$

де $t_{0.025,n-2}$ - 0.025 квантиль t-розподілу з $n - 2$ ступенями свободи.

s_{b_1} є стандартною похибкою $E(\beta_1)$,що кількісно визначає мінливість $E(\beta_1)$ обчислюється наступним чином:

$$s_{b_1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x})^2}}$$

Важливо зазначити, що менші значення s_{b_1} вказують на меншу варіабельність $E(\beta_1)$, а отже, на більшу впевненість у точності цієї оцінки. І навпаки, більші значення s_{b_1} вказують на більшу мінливість $E(\beta_1)$ і меншу впевненість у її точності.

Для оцінки сили лінійного зв'язку між x та y ми використовуємо коефіцієнт детермінації, який позначається як R^2. Він обчислюється наступним чином:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} є сумою квадратів різниць між фактичними значеннями y_i та прогнозованими значеннями $\hat{y}_i$. SS_{tot} є сумою квадратів різниць між фактичним y_i та середнім значенням $\bar{y}$.

Отже, якщо $R^2 = 1$, це означає, що модель пояснює всю варіабельність даних навколо свого середнього значення. Іншими словами, модель ідеально відповідає даним. З іншого боку, якщо $R^2 = 0$, то модель не пояснює жодної варіації даних навколо свого середнього значення.

Поліноміальна регресія [6] [7] - це статистичний метод, за допомогою якого моделюється зв'язок між незалежною змінною та залежною змінною у вигляді полінома n-го степеня.

Загальний вигляд поліноміальної регресії має наступний вигляд:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_n x_i^n + \epsilon_i$$

y_i - залежна змінна, x_i - незалежна змінна, яку ми використовуємо для передбачення залежної змінної. Кожне значення x_i підноситься до степеня від 1 до n, де n - степінь полінома. Степінь полінома визначає кількість горбів, які матиме лінія найкращої апроксимації на графіку.

Коефіцієнти β $(\beta_0, \beta_1, ..., \beta_n)$ - параметри поліноміальної регресії, які ми хочемо оцінити. Член β_0 - передбачене значення y, коли всі x_i дорівнюють нулю. ϵ_i є похибкою для i-го спостереження.

Методом найменших квадратів можна знайти значення коефіцієнтів β, які мінімізують залишкову суму квадратів (RSS), що є сумою квадратів різниць між спостереженими та передбаченими значеннями y. Значення RSS обраховується формулою:

$$RSS = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_n x_i^n))^2$$

Мінімізація виконується шляхом взяття похідної від RSS по кожному β_j, прирівнявши їх до нуля і розв'язавши рівняння для β_j. В результаті отримуємо систему лінійних рівнянь:

$$\frac{\partial RSS}{\partial \beta_j} = -2\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_n x_i^n))x_i^j = 0 \quad \forall j = 0,1,...,n$$

Кожне рівняння відповідає різному β_j, їх розв'язання дають значення β_j, які мінімізують RSS, що і є оцінкою β_j, що далі буде позначатися як $E(\beta_j)$.

Після отримання оцінок β_j важливо оцінити їх статистичну значущість. Це робиться шляхом формулювання нульової гіпотези про те, що кожен істинний коефіцієнт дорівнює нулю, що означає, що відповідна незалежна змінна не впливає на залежну змінну. Альтернативна гіпотеза полягає в тому, що коефіцієнт не дорівнює нулю, а це означає, що незалежна змінна робить свій внесок.

Значущість оціненого коефіцієнта $E(\beta_j)$ перевіряється за допомогою його t-статистики, яка є відношенням оцінки до її стандартної похибки:

$$t_j = \frac{E(\beta_j)}{SE(E(\beta_j))}$$

Тут $SE(E(\beta_j))$ - це стандартна похибка оцінки коефіцієнта $E(\beta_j)$.

Вона вимірює середню величину, на яку ця оцінка $E(\beta_j)$ коливається навколо фактичного значення β_j.

Нульовою гіпотезою для кожного t-тесту є $H_0 : \beta_j = 0$, а альтернативною - $H_1 : \beta_j \neq 0$. Якщо абсолютне значення t-статистики більше за критичне значення з таблиці t розподілу, то ми відкидаємо нульову гіпотезу і робимо висновок, що незалежна змінна дійсно суттєво впливає на залежну змінну.

Поліноміальна регресія є потужним інструментом для моделювання нелінійних зв'язків.

Регресія Лассо [8] - статистичний метод, який використовує згортання. Регресія Лассо використовується для розуміння зв'язку між залежною змінною та однією або кількома незалежними змінними, подібно до інших регресійних моделей. Унікальним аспектом регресії Лассо є її здатність виконувати як відбір змінних, так і регуляризацію з метою підвищення точності прогнозування.

Оцінки Лассо $\hat{\beta}^{lasso}$ визначаються як:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

У цьому рівнянні доданок $\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ є RSS, яке Лассо намагається мінімізувати. Член $\lambda \sum_{j=1}^{p} |\beta_j|$ є штрафним членом. λ - це параметр налаштування, який визначає силу штрафу.

Лассо обмежує суму абсолютних значень коефіцієнтів, щоб вона була меншою за певне значення. Це обмеження призводить до того, що деякі оцінки коефіцієнтів точно дорівнюють нулю, коли параметр налаштування λ є достатньо великим. Це призводить до того, що модель стає простішою і легшою для інтерпретації.

Штрафний параметр λ контролює величину згортання: чим більше значення λ, тим більша величина згортання. Значення λ можна вибрати за допомогою перехресної перевірки. Зазвичай виконується 5- або

10-кратна перехресна перевірка, і вибирається значення λ, яке призводить до найменшої перехресної середньоквадратичної помилки.

Після того, як ми отримали коефіцієнти $\hat{\beta}_j$, можна обчислити $\hat{y}$ за допомогою лінійного рівняння:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

Ефективність регресійної моделі Лассо можна оцінити за допомогою метрик, подібних до тих, що використовуються для звичайної регресії за методом найменших квадратів.

Лассо може бути особливо корисним при роботі з набором даних з великою кількістю ознак, коли потрібен відбір ознак, або коли в даних присутня мультиколінеарність.

Дерева з градієнтним підсиленням [9] - це потужний гнучкий метод для задач регресії. Це тип ансамблевого методу навчання, де нові моделі додаються для виправлення помилок, зроблених існуючими моделями. Моделі додаються послідовно до тих пір, поки не буде зроблено жодних подальших покращень. Метою дерева з градієнтним підсиленням є пошук найкращої функції $F(x)$, яка відображає вхідні змінні x у неперервну вихідну змінну y.

У контексті задачі регресії метою є пошук моделі $F(x)$, яка мінімізує математичне сподівання заданої функції втрат $L(y, F(x))$:

$$\underset{F}{\operatorname{argmin}} E_{y,x}[L(y,F(x))]$$

На практиці ми не знаємо розподілу y та x і тому не можемо обчислити математичне сподівання. Тому ми використовуємо емпіричну мінімізацію ризику і замінюємо математичне сподівання середнім значенням по навчальній вибірці:

$$\underset{F}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i,F(x_i))$$

На кожній ітерації ми застосовуємо слабкого учня для передбачення залишків (від'ємний градієнт функції втрат) попередньої моделі. Кожен слабкий навчальний елемент генерує слабке передбачення $h(x)$. Воно додається до поточної моделі, щоб мінімізувати функцію втрат:

$$F_m(x) = F_{m-1}(x) + \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i))$$

Ця процедура виконується ітераційно для покращення точності прогнозів за рахунок зменшення залишків функції втрат.

Прогнозованим значенням $\hat{y}$ моделі з M деревами є сума прогнозів кожного дерева:

$$\hat{y} = F_M(x) = \sum_{m=1}^{M} h_m(x)$$

де $h_m(x)$ - прогноз m-го дерева.

Продуктивність градієнтного підсилення сильно залежить від налаштувань його параметрів. Наприклад, швидкість навчання η контролює вплив кожного дерева на кінцевий результат. Зазвичай перевага надається меншим значенням, оскільки це робить модель стійкою до специфічних характеристик кожного дерева, що дозволяє їй добре узагальнювати невидимі дані.

Кількість дерев M та глибина дерев контролюють складність моделі. Збільшення цих значень робить модель складнішою, що потенційно може призвести до перенавчання.

Метрики оцінки регресії включають середню квадратичну похибку (MSE), кореневу середню квадратичну похибку (RMSE) та середню абсолютну похибку (MAE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i$$

Дерева з градієнтним підсиленням досягає успіху завдяки використанню концепції "навчання на помилках послідовно залучаючи слабких учнів для підвищення ефективності прогнозування.

Дерева рішень [10] - метод класифікації, що розбиває простір ознак на регіони за допомогою ієрархічних методів, заснованих на правилах, а не підганяє лінійне рівняння під спостережувані дані. Дерева рішень приймають рішення, які візуалізуються через деревоподібну модель. Дерево починається з кореневого вузла на вершині, за яким слідують нащадки, кожен з яких представляє тест на прийняття рішення щодо ознаки. Нарешті, дерево закінчується вузлами, які представляють мітки класів.

Формування дерева рішень базується на рекурсивному підході до розбиття. Вектор ознак задається як $x \in R^d$ і цільова змінна y. Дерево рішень розбиває простір ознак на M окремих областей, $R_1, R_2, ..., R_M$ і для кожної області R_j присвоює мітку класу c_j.

Це розбиття виконується на основі послідовності тестів. Тестовий вузол дерева являє собою тест на одну з ознак x_i, наприклад, $x_i < t$ для деякого t. Результат тесту визначає гілку, по якій слід рухатись, і яка веде або до іншого тестового вузла, або до листкового вузла. Кожна листкова вершина представляє область R_j простору ознак і асоціюється з міткою класу c_j.

Метою цього процесу є створення дерева, яке мінімізує ймовірність помилкової класифікації, що визначається як

$$L(T) = \sum_{j=1}^{M}\sum_{i \in R_j} I(y_i \neq c_j)$$

Тут $I(y_i \neq c_j)$ - індикаторна функція, яка дорівнює 1, якщо $y_i \neq c_j$ і 0 у протилежному випадку, а T - дерево.

Будуючи дерево рішень, ми, по суті, ставимо серію запитань, призначених для уточнення класифікації. Кожне питання, представлене вузлом у дереві, певним чином розділяє дані. Щоб вибрати, яке питання поставити, нам потрібен спосіб оцінити, наскільки якісним буде цей розподіл. Саме тут на допомогу приходять такі показники, як домішка Джині та ентропія.

Домішка Джині - це міра неправильної класифікації, яка застосовується в контексті багатокласового класифікатора. Формула має вигляд:

$$Ginni(p) = 1 - \sum_{i=1}^{J} p_i^2$$

Тут p_i - це ймовірність того, що елемент буде віднесено до класу i. Вона коливається між 0 і 1, де 0 позначає чистий вузол (всі елементи належать до одного класу), а 1 - нечистий вузол (елементи випадковим чином розподілені між різними класами).

Ентропія є мірою невпорядкованості або невизначеності. Вона забезпечує спосіб вимірювання домішок або невпорядкованості в множині. У випадку багатокласовості ентропія має таку формулу:

$$Entropy(p) = -\sum_{i=1}^{J} p_i \log_2(p_i)$$

Тут, як і раніше, p_i - це ймовірність того, що елемент буде віднесено до класу i.

Як і домішка Джині, ентропія змінюється від 0 до 1. Менша ентропія відповідає меншій невпорядкованості, що означає, що зразки у вузлі належать до одного класу. Вузол є чистим (однорідним), коли ентропія дорівнює нулю.

Отже, класифікація за допомогою дерева рішень є потужним

інструментом для вирішення проблем класифікації завдяки своїй простій інтерпретації, універсальності в роботі як з числовими, так і з категоріальними даними, а також здатності вловлювати складні, нелінійні зв'язки в даних.

Машина опорних векторів [10] - статистичний метод, який використовується для класифікації, основним принципом машини опорних векторів є створення гіперплощин, які чітко класифікують точки даних у багатовимірному просторі, без будь-якого перекриття. Мета машини опорних векторів - знайти гіперплощину в N-вимірному просторі, яка чітко класифікує точки даних.

Гіперплощина може бути описана наступним чином:

$$\vec{w} \cdot \vec{x} - b = 0$$

де $\vec{w}$ - нормальний вектор до гіперплощини, $\vec{x}$ - вхідний вектор і b - зсув.

Метою машини опорних векторів є максимізація відстані від цієї гіперплощини до найближчої точки даних будь-якого класу. Цю відстань часто називають маржею. Математичне формулювання задачі має вигляд:

$$\min_{w,b} \frac{1}{2}||w||^2 \text{ if } y_i(w \cdot x_i - b) \geqslant 1, \forall i$$

Тут y_i є міткою класу точки даних x_i і набуває значень у діапазоні $\{-1, 1\}$.

Точки даних, які лежать найближче до границі рішення, називаються опорними векторами. Лише ці вектори використовуються для обчислення маржі і функції рішення, звідси і назва "Машина опорних векторах".

Машина опорних векторів здатна виконувати нелінійну класифікацію. Це досягається за допомогою техніки, яка називається "трюк ядра". Він полягає у відображенні вхідних ознак у простір вищої

розмірності, де гіперплощина може бути використана для розділення класів.

Найпоширенішим ядром є ядро з радіально-базисною функцією:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

де $K(x, x')$ - функція ядра, x і x' - вхідні вектори, а σ - вільний параметр.

Функція прийняття рішення в машині опорних векторів, також відома як дискримінантна функція, відіграє вирішальну роль у визначенні поділу між класами.

Функція рішення має вигляд:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

де α_i - коефіцієнти, x_i - опорні вектори, K - функція ядра, n - кількість опорних векторів.

Таким чином, функція рішення є фундаментальним аспектом класифікатора, який визначає межу між класами.

Машина опорних векторів є потужним інструментом який підходить до завдання, зосереджуючись на максимізації маржі та опорних векторах. Цей підхід часто дає надійні та точні моделі.

Метод наївної класифікації Байєса [10] - заснований на правилі Байєса - фундаментальному принципі теорії ймовірностей і статистики, який описує, як оновити переконання, маючи певні докази. Іншими словами - воно описує акт навчання.

Наївний класифікатор Байєса працює на основі припущення, що кожна ознака в наборі даних є умовно незалежною від будь-якої іншої ознаки, заданої змінною класу. Це припущення, і саме воно є причиною "наївності"наївного Байєса.

Математичною основою наївного класифікатора Байєса є правило Байєса, яке формулюється наступним чином:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

.

Наївний байєсівський класифікатор робить припущення, що кожна ознака X_i у векторі ознак X є умовно незалежною від кожної іншої ознаки X_j (для $i \neq j$), заданої змінною класу Y. Це припущення дозволяє нам спростити обчислення ймовірності P(X|Y) наступним чином:

$$P(X|Y) = P(X_1, X_2, ..., X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Це рівняння показує, що ймовірність P(X|Y) можна обчислити як добуток ймовірностей P(X_i|Y) для i = 1, ..., n. Це значне спрощення, оскільки воно зводить проблему оцінки спільного розподілу ймовірностей за n змінними до проблеми оцінки n індивідуальних розподілів ймовірностей.

Для нового екземпляра $X_{new} = \langle X_1 \ ... \ X_n \rangle$ наївний байєсівський класифікатор обчислює ймовірність того, що Y прийме будь-яке задане значення y_k формула виглядає наступним чином:

$$P(Y = y_k|X) = \frac{P(Y = yk) \prod_i P(Xi|Y = yk)}{\sum_j P(Y = yj) \prod_i P(Xi|Y = yj)}$$

Це рівняння показує, як обчислити апостеріорну ймовірність P(Y|X) на основі попередньої ймовірності P(Y), ймовірності P(X|Y) та доказів P(X). Знаменник дробу є константою нормалізації, яка гарантує, що апостеріорні ймовірності дорівнюють одиниці для всіх можливих значень Y.

Незважаючи на свою простоту і наївність припущень, наївний байєсівський класифікатор виявився надійним і універсальним інструментом у багатьох застосуваннях. Він особливо ефективний у просторах високої розмірності, де кількість ознак велика порівняно з кількістю точок даних. У таких випадках наївний байєсівський

класифікатор може перевершити більш складні класифікатори, які не роблять припущення про умовну незалежність.

Отже, класифікатор наївного Байеса - це простий, але потужний інструмент у машинному навчанні та аналізі даних.

2.2 Побудовані регресії та класифікації

Наступний розділ роботи присвячений представленню та порівнянню результатів, регресійних моделей, а також класифікацій. Ефективність і точність прогнозування цих моделей буде оцінюватися за допомогою різноманітних метрик. Для ясності та інтуїтивності числових результатів зроблені візуалізації, які явно показують різницю між результатами.

Спочатку стратегія полягала в тому, щоб навчити регресійні моделі на вибірці характеристик, які, як вважалося, найбільше впливають на ціни на землю - площа, висота над рівнем моря, схил, якість землі та відстань до Києва. Однак, отримана модель, показала мале значення R^2, що свідчить про слабку прогностичну силу та відсутність достатньої кількості пояснювальних змінних. Усвідомлюючи потребу в ширшому спектрі даних, було вирішено використати всі ознаки, наявні в наборі даних. Однак, незважаючи на такий розширений підхід, ефективність моделей можна було б покращити. Отже методологія була покращена, перетворивши безперервний результат на категоріальний. Ціни були розділені на сім окремих категорій - "зі знижкою" , "дуже дешево" , "дешево" , "середньо" , "дорого" , "дуже дорого" і "преміум". Для забезпечення комплексного аналізу ми провели чотири окремі оцінки для кожної моделі. Перша оцінка передбачала моделювання цін на землю по всій країні, друга оцінка ґрунтувалася на першій шляхом включення цінових категорій, третя оцінка моделювала ціни на землю виключно в Києві, четверта ґрунтувалася на третій шляхом включення цінових категорій.

Проведемо аналіз результатів, коли кожна модель навчалася та оцінювалася з використанням одного і того ж набору даних по всій країні, без будь-якої категоризації на основі цінового діапазону.

Як показано в таблиці 2.1, жодна з моделей не дала гарних результатів. Це підкреслює складність набору даних і проблеми, пов'язані з прогнозуванням цін на землю. Незважаючи на ретельну попередню обробку, максимальна оцінка R^2 становила лише 0,213, що свідчить про те, що значна частина дисперсії цін залишається непоясненою за допомогою цих моделей.

Таблиця 2.1 – Результати моделей по всій країні не розподілені за категоріями

Model	R^2	Mean squared error
Linear	0.213	2322019415171
Polynomial	0.183	3366246663764
LASSO	0.051	3913451328573
Gradient Boost	0.017	4051536847369

Ці результати свідчать про те, що для прогнозування цін на землю по всій країні може знадобитися більш складний підхід до моделювання.

Проведемо аналіз результатів, коли кожна модель навчалася та оцінювалася з використанням одного і того ж набору даних по всій країні, з категоризацією на основі цінового діапазону.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	174856.930	333.124	418.159	1.05e-11	0.303
Very Cheap	284110875.365	13329.476	16855.589	3.87e-10	0.235
Cheap	753905188.187	13226.455	27457.334	1.69e-09	0.235
Pricy	14228814283.250	93856.981	119284.594	-3.72e-09	0.379
Expensive	2.72e-14	1.07e-07	1.65e-07	2.56e-09	0.998
Very Expensive	1707233206264.270	933758.354	1306611.345	-9.05e-09	0.588
Premium	3.13e-09	3.29e-05	5.60e-05	-2.17e-05	0.979

Таблиця 2.2 – Лінійна регресія з розподілом на категорії по всій країні

У таблиці 2.1 для лінійнійної регресії існує різниця в результатах для різних категорій. Якщо поглянути на категою "дисконт" метрики, такі як середня квадратична похибка, середня абсолютна похибка та середньоквадратична похибка є відносно низькими порівняно з деякими іншими категоріями, що свідчить про те, що модель є доволі точною в цій категорії. Показник R^2, є відносно низьким, що свідчить про те, що модель пояснює лише близько 30 % дисперсії залежної змінної.

З іншого боку, в категоріях "дорогі" та "преміум" результат R^2 дуже близький до 1. Такий ідеальний результат може вказувати на надмірне припасування, коли модель буде працювати з іншими даними результат може бути незадовільним. Я вважаю, що це є результатом моєї попередньої обробки, оскільки для цих категорій я зробив дуже вузький ціновий діапазон та замінив пропущенні значення середнім, що теж може сприяти моделі лінійоної регресії.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	69383.195	200.201	263.407	-3.22e-11	0.724
Very Cheap	202050335.864	10842.626	14214.441	3.69e-09	0.456
Cheap	563304756.270	12430.795	23734.042	5.05e-10	0.428
Pricy	8.42e-11	4.80e-06	9.18e-06	2.81e-06	0.468
Expensive	1.88e-16	9.27e-09	1.37e-08	8.79e-10	0.747
Very Expensive	2.98e-11	4.27e-06	5.46e-06	4.08e-06	0.573
Premium	1.69e-09	3.31e-05	4.11e-05	2.93e-05	0.515

Таблиця 2.3 – Поліноміальна регресія з розподілом на категорії по всій країні

Поліноміальна регресія у таблиці 2.3, демонструє значні покращення порівняно з лінійною регресією, найкращі результати демонструє 4 ступінь поліному. Категорія "дисконт" демонструє різке збільшення показника R^2, що свідчить про те, що поліноміальна модель більше підходить для цих даних. Це може вказувати на нелінійний зв'язок між змінними та ціною.

Категорії "дуже дешево" та "дешево" також демонструють покращення показників R^2 порівняно з лінійною регресією, що свідчить про те, що ці дані також мають нелінійний зв'язок із залежною змінною.

Цікаве спостереження можна зробити для категорій "дорогі" та "дуже дорогі". В обох цих категоріях MSE, MAE та RMSE є надзвичайно низькими.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	151.400	8.964	12.304	2.87e-14	0.999
Very Cheap	5333823.121	1633.728	2309.507	-4.91e-13	0.986
Cheap	108344136.504	2925.492	10408.849	3.41e-11	0.890
Pricy	8912196833.749	75249.720	94404.432	-1.99e-11	0.611
Expensive	1234183889.167	30608.500	35130.953	-6.37e-11	0.374
Very Expensive	1462537703935.850	862932.948	1209354.251	9.75e-10	0.647
Premium	250222849109944.160	9238636.914	15818433.839	5.09e-09	0.496

Таблиця 2.4 – Модель регресії дерев з градієнтним підсиленням з розподілом на категорії по всій країні

Результати регресії з градієнтим підсиленням у таблиці 2.4 показують багатообіцяючі значення R^2 у більшості категорій, особливо в категоріях "знижка" та "дуже дешево". Ця модель відома своєю здатністю вловлювати складні закономірності, що може пояснити її успіх тут. Для тренування моделі були використані різні параметри швидкості навчання від 0.1 до 0.001, максимальної глибини від 2 до 20 та n оцінювачей від 100 до 500.

Метрики похибки є нижчими в категоріях "знижка" та "дуже дешево" порівняно з попередніми моделями, що вказує на те, що модель регресії з градієнтим підсиленням точно прогнозує ціну в цих категоріях.

Навпаки, в категорії "дорогі" оцінка R^2 є нижчою, а MSE та RMSE є значно вищими, ніж у попередніх моделей, що свідчить про те, що ця модель може бути непридатною для цієї категорії.

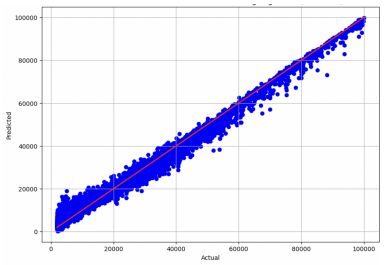
Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	217791.980	384.340	466.682	1.33e-13	0.132
Very Cheap	348692260.282	15076.272	18673.303	1.24e-12	0.061
Cheap	891785319.970	13609.841	29862.775	2.98e-11	0.095
Pricy	21434315138.141	123375.085	146404.628	-4.64e-11	0.065
Expensive	1706769988.530	34795.082	41313.073	-7.60e-11	0.135
Very Expensive	3172537926202.163	1300876.375	1781161.960	1.25e-09	0.235
Premium	383043041706300.800	11141086.415	19571485.424	3.48e-09	0.229

Таблиця 2.5 – Модель Лассо з розподілом на категорії по всій країні

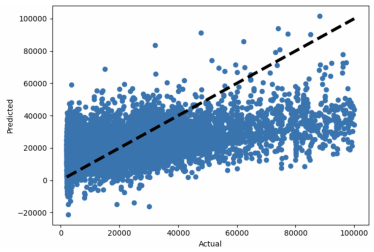
Регресійна модель Лассо, результати якої наведено у таблиці 2.5 демонструє нижчі показникими R^2 та вищі показники похибки порівняно з іншими моделями. Це може свідчити про те, що лінійні припущення моделі Лассо можуть бути недостатніми для врахування складності даних. Були спробувані різні значення альфа моделі, але на жаль, значних покращень, не було отримано.

Важливим спостереженням тут є високі показники похибки в категорії "дуже дорогі".

Нижче наведено цікаві візуалізації на рисунках 2.1(а), 2.1(б), 2.2(а), 2.2(б) що стосуються цих вимірювань:

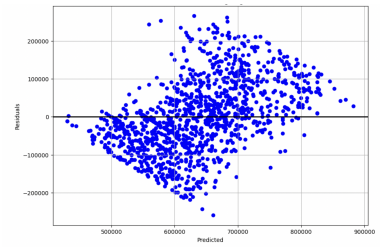


(а) Дешева категорія для моделі дерев з градієнтним підсиленням

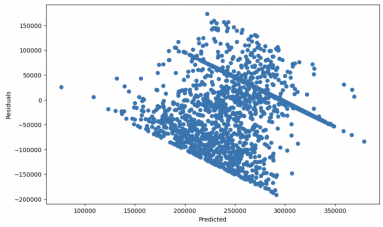


(б) Лінійна регресія дуже дешева категорія

Рисунок 2.1 – Діаграми розсіювання



(а) Категорія середньої ціни для моделі дерев з градієнтним підсиленням



(б) Поліноміальна регресія дешева категорія

Рисунок 2.2 – Залишкові ділянки

Мені стало цікаво чи можна покращити результати, якщо сконцентруватися на даних у конкретному місті, оскільки значення сильно відрізняються в різних регіонах країни.

Таблиця 2.6 – Результати моделей для Києва не розподілені за категоріями

Model	Mean Squared Error	R^2
Linear	11692355751298.607	0.1391975756667776
Polynomial	11692355751298	0.1391975756
LASSO	6495764005976	0.048761489
Gradient Boost	4062373292634.193	0.015133888450460709

Порівнюючи ці результати у таблиці 2.6 з попереднім аналізом на основі даних по всій країні, бачимо, що результати для всієї країни були краще. Ефективність моделі зазвичай залежить від кількості та якості наявних даних. Київ, будучи лише частиною країни, надає менший набір даних.

Проведемо аналіз результатів, коли кожна модель навчалася та оцінювалася з використанням одного і того ж набору даних для Києва, з категоризацією на основі цінового діапазону.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	4.38e-22	1.87e-11	2.09e-11	1.81e-11	0.999
Very Cheap	2.15e8	11304	14651	0.13	0.698
Cheap	3.56e8	14975	18859	-23.32	0.943
Pricy	7.95e9	70924	89162	16.73	0.663
Expensive	2.31e8	9690	15190	-0.20	0.559
Very Expensive	4.73e11	486040	687581	-124.10	0.849
Premium	9.31e-14	2.42e-07	3.05e-07	1.92e-07	0.999

Таблиця 2.7 – Лінійна регресія з розподілом на категорії для Києва

Для категорії "дисконт" у таблиці 2.7 модель демонструє надзвичайну точність з надзвичайно низькими показниками похибок.

На противагу цьому, категорії "дуже дешево" , "середньо" , "дорого" і "дуже дорого" , демонструють значно вищі показники похибки, а також нижчі показники R^2. Ці показники свідчать про те, що лінійна модель має простір для вдосконалення при прогнозуванні в межах цих категорій.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	2.14e-22	1.04e-11	1.46e-11	8.37e-12	1.0
Very Cheap	2.26e8	11608	15023	0.04	0.683
Cheap	9.32e8	23604	30521	-9.67	0.851
Pricy	1.28e10	89849	112921	2.62	0.460
Expensive	2.46e8	10023	15694	-0.56	0.529
Very Expensive	5.77e11	536541	759433	-22.83	0.815
Premium	1.09e-13	2.98e-07	3.30e-07	2.91e-07	1.0

Таблиця 2.8 – Поліноміальна регресія з розподілом на категорії для Києва

Поліноміальна регресія у таблиці 2.8, має кращі результати порівнюючи з лінійною регресією. Категорія "дисконт" демонструє вражаючу оцінку R^2 у 1.0, що означає, що поліноміальна модель може більш ефективно відображати зв'язок між предикторами та залежними змінними в цій категорії.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	4.92e2	17.06	22.17	-3.65e-14	0.998
Very Cheap	4.41e7	5185	6643	9.61e-13	0.938
Cheap	3.01e8	12899	17335	1.89e-11	0.952
Pricy	1.08e10	86613	104038	-2.50e-10	0.541
Expensive	3.92e7	2461	6264	9.47e-11	0.925
Very Expensive	5.32e11	529747	729105	1.41e-11	0.830
Premium	1.44e14	8.98e6	1.20e7	5.25e-09	0.684

Таблиця 2.9 – Модель регресії дерев з градієнтним підсиленням з розподілом на категорії для Києва

Регресія дерев з градієнтним підсиленням у таблиці 2.9 демонструє багатообіцяючі значення R^2 у більшості категорій, особливо в категоріях "знижка" та "дуже дешево".

Крім того, показники похибки у цих двох категоріях помітно нижчі, ніж у попередніх моделях, що свідчить про вищий рівень точності. З

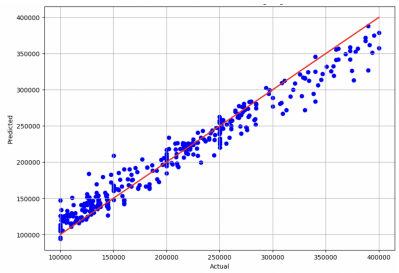
іншого боку, "середньо" категорія має нижчий показник R^2 у поєднанні з вищими показниками MSE та RMSE, ніж у попередніх моделях. Цей контраст свідчить про те, що модель регресії дерев з градієнтним підсиленням може не підходити для цієї категорії так ефективно, як для інших.

Category	**MSE**	**MAE**	**RMSE**	**Mean Signed Difference**	R^2
Discount	3.84e4	143	196	1.10e-13	0.824
Very Cheap	5.59e8	20176	23651	1.73e-12	0.213
Cheap	4.59e9	52944	67747	-1.25e-12	0.267
Pricy	2.08e10	120554	144317	6.21e-11	0.118
Expensive	4.38e8	11276	20922	1.11e-10	0.163
Very Expensive	2.60e12	1.14e6	1.61e6	-8.49e-11	0.168
Premium	2.44e14	1.03e7	1.56e7	-1.60e-07	0.463

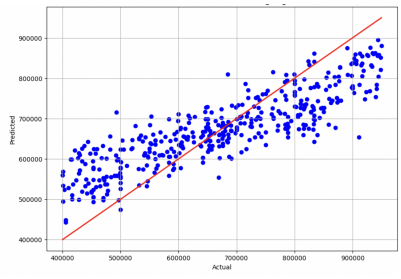
Таблиця 2.10 – Модель Лассо з розподілом на категорії для Києва

Регресія Лассо, результати якої наведено у таблиці 2.10 демонструє нижчі показники R^2 та вищі показники похибки порівняно з іншими моделями. Ця узгодженість означає, що лінійні припущення Лассо можуть недостатньо відображати складність даних.

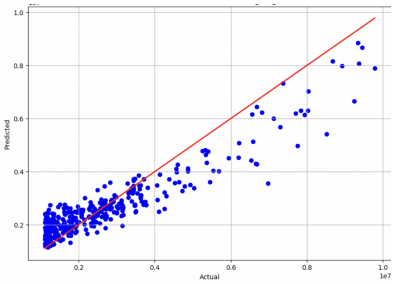
Я вважаю, що найкращі результати демонструє модель дерев з градієнтним підсиленням, отже демонструю її візуалізації на рисунках 2.3(а), 2.3(б), 2.3(в), 2.4(а), 2.4(б), 2.4(в) :



(а) Дешева категорія

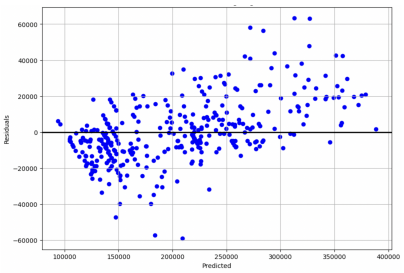


(б) Категорія середньої ціни

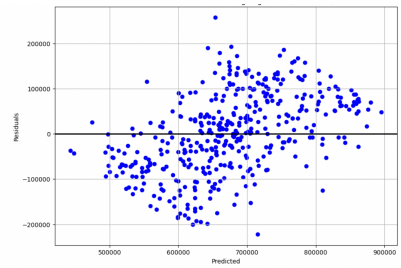


(в) Дуже дорога категорія

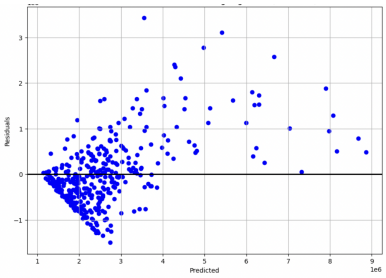
Рисунок 2.3 – Діаграми розсіювання



(а) Дешева категорія



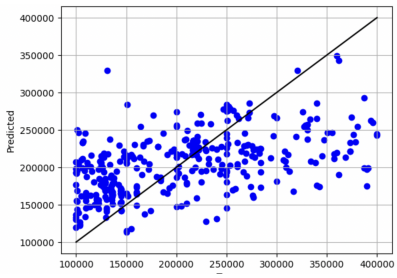
(б) Категорія середньої ціни



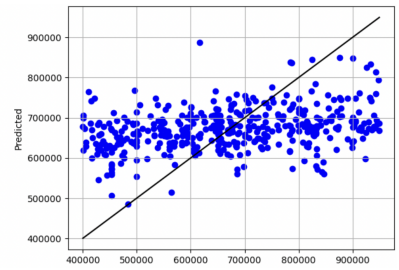
(в) Дуже дорога категорія

Рисунок 2.4 – Залишкові ділянки

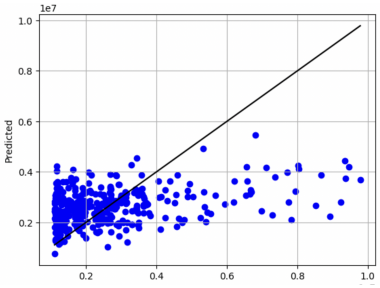
Для кращого розуміння, на скільки добре впоралась з поставленою задачею модель дерев з градієнтним підсиленням продемонстую декілька візуалізацій моделі Лассо на рисунках 2.5(а), 2.5(б), 2.5(в), яка продемонструвала найгірші результати:



(а) Дешева категорія



(б) Категорія середньої ціни



(в) Дуже дорога категорія

Рисунок 2.5 – Діаграми розсіювання

Щоб заглибитися в тему дослідження якості землі мені стало цікаво класифікувати такі параметри, як основана вирощувана культура, регіони та тип власності різними методами та подивитися на результати.

Нижче продемонстровані результати у таблицях 2.11, 2.12, 2.13 різних моделей для класифікації основної вирощуваної культури:

Accuracy	Cohen's kappa
0.909	0.877

Таблиця 2.11 – Дерева рішень

Accuracy	Cohen's kappa
0.593	0.515

Таблиця 2.12 – Метод наївної класифікації Байєса

Accuracy	Cohen's kappa
0.251	0.139

Таблиця 2.13 – Машина опорних векторів

Найкращі результати демонструє метод дерев рішень, для використаного датасету він побудував 1817 тестів, на жаль, візуалізація занадто велика, отже я покажу перші та останні два тести у вигляді таблиці 2.14 :

Rule	Record	Count	Number of correct
1	LC_{2018} <= 1.5 AND LC_{2016} <= 3.5 AND LC_{2017} <= 9.0 AND LC_{2017} <= 10.5 => "1.0"	339	338
2	LC_{2021} <= 2.5 AND LC_{2021} <= 4.0 AND LC_{2016} <= 2.5 AND LC_{2019} <= 4.0 AND LC_{2018} > 1.5 AND LC_{2016} <= 3.5 AND LC_{2017} <= 9.0 AND LC_{2017} <= 10.5 => "2.0"	1240	1240
1816	LC_{2021} > 12.5 AND LC_{2016} <= 13.5 AND LC_{2019} > 11.5 AND LC_{2019} > 10.0 AND LC_{2016} > 11.5 AND LC_{2016} > 10.5 AND LC_{2016} > 9.5 AND LC_{2017} > 10.5 => "13.0"	3	3
1817	LC_{2016} > 13.5 AND LC_{2019} > 11.5 AND LC_{2019} > 10.0 AND LC_{2016} > 11.5 AND LC_{2016} > 10.5 AND LC_{2016} > 9.5 AND LC_{2017} > 10.5 => "14.0"	12	11

Таблиця 2.14 – Дерево рішень по основним вирощуманим культурам

Нижче продемонстровані результати у таблицях 2.15, 2.16, 2.17 різних моделей для класифікації регіону:

Accuracy	Cohen's kappa
0.963	0.961

Таблиця 2.15 – Дерева рішень

Accuracy	Cohen's kappa
0.663	0.644

Таблиця 2.16 – Метод наївної класифікації Байєса

Accuracy	Cohen's kappa
0	0

Таблиця 2.17 – Машина опорних векторів

Метод опрних векторів зовсім не підходить для класифікації регіонів, а метод наївного Байєса, трохи покращив результати порівняно с класифікацією основної вирощуваної культри, але знову найкращі результати демонструє метод дерев рішень, для використаного датасету він побудував 944 теста, продемонстровано перші та останні два тести у вигляді таблиці 2.18 :

Rule	Record	Count	Number of correct
1	$Dist\_Citie$ <= 0.022 AND $Kyiv\_dist$ <= 59.673 AND $Kyiv\_dist$ <= 87.14850000000001 => "32"	112	111
2	$Bonitet$ <= 52.5 AND $Kyiv\_dist$ <= 24.933 AND $Dist\_PR\_Ro$ <= 16.286 AND $Kyiv\_dist$ > 0.0015 AND $Kyiv\_dist$ <= 59.673 AND $Kyiv\_dist$ <= 87.14850000000001 => "18"	4	4
943	$BonitetL$ > 56.0 AND $Kyiv\_dist$ > 351.28200000000004 AND $BonitetL$ <= 62.0 AND $Altitude$ > 169.7730499 AND $BonitetL$ > 53.5 AND $BonitetL$ > 47.5 AND $Kyiv\_dist$ > 310.1735 AND $Kyiv\_dist$ > 87.14850000000001 => "26"	14	13
944	$BonitetL$ > 62.0 AND $Altitude$ > 169.7730499 AND $BonitetL$ > 53.5 AND $BonitetL$ > 47.5 AND $Kyiv\_dist$ > 310.1735 AND $Kyiv\_dist$ > 87.14850000000001 => "73"	164	162

Таблиця 2.18 – Дерева рішень по регіонам

Нижче продемонстровані результати у таблицях 2.19, 2.20, 2.21 різних моделей для класифікації на основі типу власності:

Accuracy	Cohen's kappa
0.901	0.664

Таблиця 2.19 – Дерева рішень

Accuracy	Cohen's kappa
0.764	0.212

Таблиця 2.20 – Метод наївної класифікації Байєса

Accuracy	Cohen's kappa
0	0

Таблиця 2.21 – Машина опорних векторів

Метод опорних векторів зовсім не підходить для класифікації типу власності, а метод наївного Байєса, трохи покращив результати порівняно з минулими класифікаціями, але знову найкращі результати демонструє метод дерев рішень, для використаного датасету він побудував 1588 теста, продемонстровано перші та останні два тести у вигляді таблиці 2.22 :

Rule	Record	Count	Number of correct
1	$priceha$ <= 3777.6333495 AND $Dist\_Citie$ <= 3.8834999999999997 AND $area$ <= 1.994599998 AND $oblast$ = "63"AND $priceha$ <= 295115.27262318577 AND $area$ <= 2.0072499515 => "Commercial"	17	16
2	LC_2021 <= 11.5 AND $Dist\_COUN\_$ <= 0.0795 AND $Dist\_PR\_Ro$ <= 1.8405 AND $Altitude$ <= 145.8143135 AND $priceha$ > 3777.6333495 AND $Dist\_Citie$ <= 3.8834999999999997 AND $area$ <= 1.994599998 AND $oblast$ = "63"AND $priceha$ <= 295115.27262318577 AND $area$ <= 2.0072499515 => "Personal"	6	6
1587	$oblast$ = "44"AND $area$ > 2.0072499515 => "Commercial"	508	508
1588	$oblast$ = "21"AND $area$ > 2.0072499515 => "Commercial"	25	25

Таблиця 2.22 – Дерева рішень по типу власності

Під час аналізу метод дерева рішень постійно виділявся, забезпечуючи вражаючі результати. Сила моделі дерева рішень полягає в її здатності обробляти великі масиви даних і складні взаємозв'язки, що було очевидно для нашого набору даних.

Висновки до розділу 2

Головна мета полягала в тому, щоб знайти найкращу модель для опису даних. Метою було не лише перевірити можливості цих моделей, але й визначити найкращу комбінацію параметрів, яка дає оптимальні результати. Щоб підвищити точність моделей, було вирішено розділити дані на основі цінового діапазону. Цей підхід дозволив значно покращити результати всіх моделей, що свідчить про важливість стратегічного підходу до обробки даних для ефективності моделей. Щоб підвищити точність був проведений аналіз конкретно для Києва. Ця спеціальна стратегія виявилася успішною, надаючи моделям більш детальне розуміння регіональної специфіки. Серед усіх моделей найкращі результати продемонструвала модель дерев з градієнтним підсиленням. Сильна сторона моделі полягає в її здатності обробляти складні структури даних і відображати складні взаємозв'язки між змінними. Далі, щоб отримати всебічне розуміння українських земель, ми застосували класифікаційні моделі для аналізу різних параметрів. Модель дерева рішень виділилася завдяки своїй винятковій здатності знаходити складні взаємозв'язки в даних.

ВИСНОВКИ

Основною метою роботи було використання математичних методів для кращого розуміння динаміки землекористування в Україні. Спочатку були провдені статистичні виміри та тести, щоб зрозуміти загальний стан земель в Україні та виявити наявність або відсутність взаємозв'язків між різними параметрами в наборі даних. Ці перші кроки мали вирішальне значення для формування нашого розуміння набору даних. Важлвим етапом була попередня обробка даних перед побудовою моделей. Після цього були побудовані перші версії регресійних моделей. Спочатку зусилля не дали очікуваних результатів. Для покращення результатів дані були розділені на основі цінового діапазону, що призведе до покращення продуктивності моделі. Гіпотеза виявилася правильною, і цей підхід значно підвищив точність усіх наших моделей. Це підкреслює важливість стратегічного підходу до обробки даних. З метою подальшого покращення точності ми провели поглиблений аналіз, специфічний для Київської області. Ця регіональна стратегія забезпечила нашим моделям більш детальне розуміння структури землекористування в Києві, і вона виявилася успішною в підвищенні продуктивності моделей. Серед усіх моделей, які ми оцінювали, найкращою виявилася модель дерев з градієнтним підсиленням. Її здатність обробляти складні структури даних і відображати складні взаємозв'язки в наших даних була безпрецедентною. Вийшовши за рамки регресійного аналізу, ми також застосували моделі класифікації, щоб отримати всебічне розуміння українських земель. Знову ж таки, модель дерева рішень виділилася завдяки своїй винятковій здатності обробити складні взаємозв'язки в даних. Для подальших досліджень буде зібрано більше даних, та прикладено більше зусиль для їх попередньої обробки, оскільки це є ключовим фактором для успіху. Також будуть протестовані інші моделі та більша кількість параметрів.

ПЕРЕЛІК ПОСИЛАНЬ

1. Statistics in a nutshell by Sara Boslaf.

2. LibreTexts Statistics [Електронний ресурс]. — Режим доступу: https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/04%3A_Expected_Value/4.04%3A_Skewness_and_Kurtosis.

3. LibreTexts Statistics [Електронний ресурс]. — Режим доступу: https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Quantitative_Research_Methods_for_Political_Science_Public_Policy_and_Public_Administration_(Jenkins-Smith_et_al.)/06%3A_Association_of_Variables/6.01%3A_Cross-Tabulation.

4. Scribbr [Електронний ресурс]. — Режим доступу: https://www.scribbr.com/statistics/chi-square-test-of-independence/.

5. Regression analysis by example Samprit Chatterjee and Ali S. Hadi.

6. Polynomial regression [Електронний ресурс]. — Режим доступу: http://home.iitk.ac.in/~shalab/regression/Chapter12-Regression-PolynomialRegression.pdf.

7. Polynomial regression [Електронний ресурс]. — Режим доступу: https://drs.icar.gov.in/Electronic-Book/module5/11PRegres.pdf.

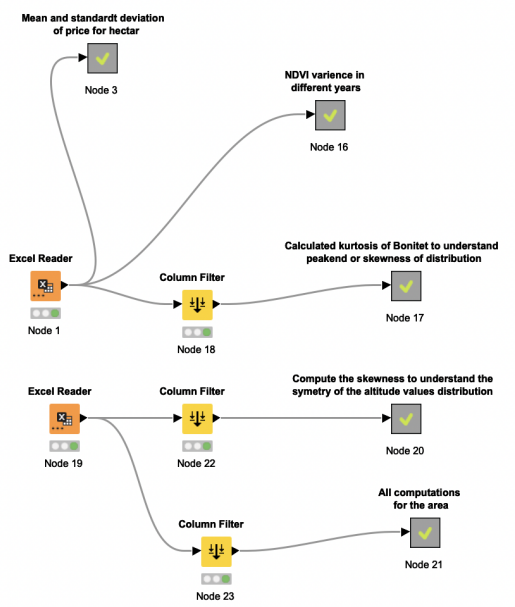
8. Statistical Learning with Sparsity by Trevor Hastie, Robert Tibshirani, and Martin Wainwright.

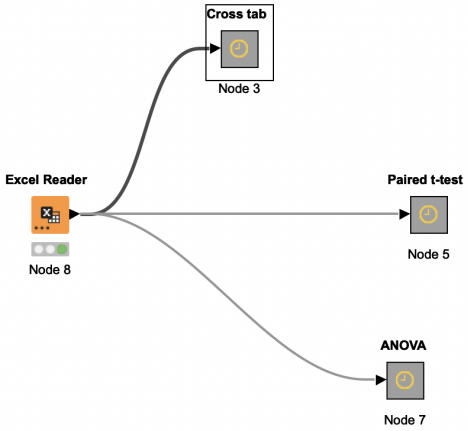
9. Hands-On Gradient Boosting by Corey Wade and Kevin Glynn.

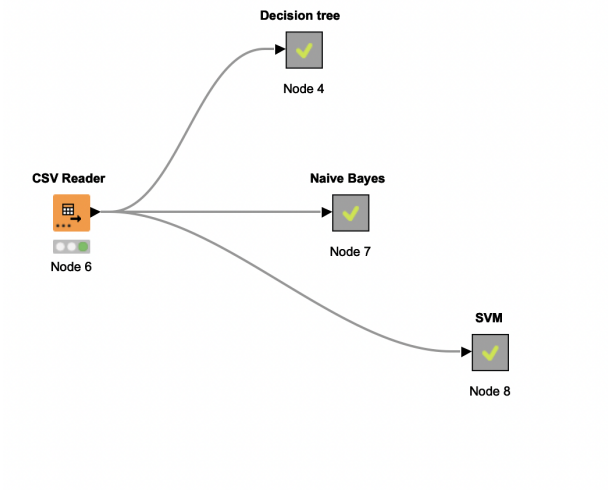
10. Data science from scratch [Електронний ресурс]. — Режим доступу: https://www.m-fozouni.ir/wp-content/uploads/2020/08/Joel_Grus_Data_Science_from_Scratch_First_Princ.pdf.

ДОДАТОК А ТЕКСТИ ПРОГРАМ

На жаль, я не маю ліцензії Knime hub, отже не можу вивантажити воркфлоу на сервер, я надаю скріншоти своїх конструкторів:







Код регресійних моделей завантажений на гітхаб